OM SHARMA

AI Engineer | Real-Time AI Systems | Voice + LLM Infrastructure
Bangalore, India • +91 8178980973 • justomsharma@gmail.com
linkedin.com/in/omsharmaofficial • github.com/justomsharma • medium.com/@OmsharmaOfficial

SUMMARY

Software engineer building production AI systems with a focus on real-time inference, streaming voice pipelines, and LLM orchestration. Shipped a voice AI interviewer handling 300+ daily conversations at sub-2-second p50 latency. Two years designing distributed backends, fine-tuning LLMs (LoRA/QLoRA), and deploying ML applications on AWS and Azure. Interested in inference optimization, speculative decoding, and small-model routing for latency-bound systems.

EXPERIENCE

Jobtwine — Associate Software Engineer *Bangalore, India | Jul 2025 – Present*

- Architected a real-time AI interviewer with a streaming STT → LLM → TTS pipeline at sub-2-second p50 latency, used by enterprise clients including Meesho, Brillio, and Deutsche Bank.
- Implemented streaming orchestration over Twilio Programmable Voice (inbound and outbound calls) bridged to LiveKit (WebRTC/SIP) with token-level TTS streaming and endpoint prediction, reducing per-turn latency by ~3 seconds.
- Built queue-based dispatch with fallback across multiple LLM providers, supporting peak loads of 300+ interviews/day with reliability under provider failures.
- Developed FastAPI backend services with PostgreSQL, Redis, and MySQL for scheduling, interview playbooks, and recruiter feedback workflows.
- Deployed and maintained production infrastructure on Azure and LiveKit Cloud as part of a 5-engineer team.

Darwix AI — Software Engineer, AI Systems *Gurugram, India | May 2025 – Jul 2025*

- Built a scalable document ingestion engine using Pinecone and PostgreSQL with automated document-type detection, improving chunking accuracy by 40%.
- Developed a real-time sales call analysis platform with speaker diarization, live transcription, and performance scoring for coaching workflows.
- Engineered a cross-platform Windows/Linux desktop client streaming dual-channel call audio over WebSocket pipelines.
- Integrated OpenAI and Anthropic APIs for automated conversation analysis; provisioned AWS RDS and EC2 infrastructure for real-time analytics workloads.

VDOIT Technologies — Software Engineer, AI/ML *Gurugram, India | Jan 2024 – Apr 2025*

- Built backend systems supporting 100K+ concurrent users using Django and multithreaded architecture.
- Designed retrieval-augmented generation (RAG) pipelines with vector databases for contextual document search.
- Fine-tuned large language models using LoRA and QLoRA techniques for domain-specific use cases.
- Deployed AI applications on AWS (EC2, S3) reducing cloud costs 20%; built CI/CD pipelines (GitHub Actions) improving deployment speed 40%. Awarded STAR Performer.

SELECTED WRITING

Where the Milliseconds Go: Anatomy of a Sub-2s Voice AI Pipeline — *Medium, May 2026*
Technical deep-dive on latency engineering in real-time voice AI: streaming end-to-end, learned endpoint prediction, speculative LLM dispatch, first-token-aware model routing, and connection warm-pooling.

PROJECTS

JiraGenie — *Natural-language CLI for Jira · github.com/justomsharma/jiragenie*

- AI agent that manages Jira tasks via natural language, with LLM-powered intent parsing, a workflow walker that finds reachable paths through Jira's transition graph (loop-guarded), and a repair layer that auto-fills required fields and retries on 400 errors.

OncoVision — *Cancer Detection from Medical Imaging*

- Convolutional neural network for detecting cancer patterns in medical imaging datasets. Built preprocessing pipelines for image normalization and augmentation; evaluated using precision, recall, and ROC-AUC.

SKILLS

Languages: Python, JavaScript / TypeScript, Node.js, SQL
AI / ML: LLM orchestration, RAG pipelines, LoRA / QLoRA fine-tuning, LangChain, LlamaIndex, LangGraph, speech processing (STT / TTS), streaming inference, vector search, speculative decoding
Frameworks: PyTorch, Hugging Face Transformers, FastAPI, Django
Real-Time Systems: Twilio Programmable Voice, SIP telephony, LiveKit, WebRTC, WebSockets, streaming pipelines, microservices, distributed orchestration
Infrastructure: AWS (EC2, S3, RDS), Azure, Docker, GitHub Actions, CI / CD
Databases: PostgreSQL, MySQL, Redis, MongoDB, Pinecone, FAISS

EDUCATION

B.Tech, Computer Science (AI & ML Specialization) *2020 – 2024*
Dronacharya College of Engineering, Gurugram

CERTIFICATIONS & ACHIEVEMENTS

Google Cloud Certified – Associate Cloud Engineer • Winner, Dronathon Intercollegiate Hackathon • STAR Performer, VDOIT Technologies (2024) • HackerRank Gold (Python); Certifications in Python and SQL