Utkarsh Tiwari

utkarshtiwar89@gmail.com | +918923011859 | linkedin.com/in/utkarsh-tiwari | github.com/UtkarshTiwari07

Summary

Production-focused AI Engineer specializing in enterprise-grade AI voice systems, with proven expertise architecting scalable multi-tenant infrastructures processing 2,000+ concurrent calls across 13+ organizations. Demonstrated proficiency in end-to-end AI system development—from LLM integration and STT/TTS optimization to backend architecture and database performance tuning. Achieved quantifiable results: 90-95% query optimization, 66% performance improvement, 80% accuracy enhancement, and 45% latency reduction in production environments.

Experience

AI Engineer, Bigship Technologies June 2025 – Present

- Drove end-to-end development of production-scale AI voice infrastructure using Python, orchestrating 500+ concurrent intelligent calls and 5,000+ daily calls, whilst architecting advanced LLM-powered conversational AI flows with real-time STT/TTS integration—achieving 80% response accuracy improvement.
- Engineered and optimized sophisticated multi-provider telephony pipelines integrating GPT-4, Deepgram STT, and ElevenLabs TTS through custom Python microservices, resulting in 45% latency reduction and seamless agent handoff capabilities, whilst implementing intelligent voicemail detection, DTMF support, and dynamic call routing to enhance enterprise-grade voice automation scalability.

AI Engineer (Freelance), Scicom.ai October – December 2025

- Architected and deployed a production-grade realtime AI voice avatar system with Retrieval-Augmented Generation (RAG), integrating LiveKit Agents, Groq LLaMA-3.3-70B, Deepgram Aura TTS, and Beyond Presence avatars to enable real-time, enterprise-scale conversational AI for customer service workflows.
- Engineered and optimized an asynchronous RAG pipeline using Pinecone serverless vector database and Sentence Transformers embeddings, achieving 10.4x cold-start latency reduction (3.9s to 378ms), thread-safe concurrent session handling, and scalable knowledge ingestion for high-availability production deployments.

Backend AI Voice Developer (Freelance), Zudu.ai October – November 2025

- Architected enterprise-grade multi-tenant backend infrastructure integrating LiveKit with multi-provider telephony systems (Twilio, Plivo) to orchestrate 2,000+ concurrent AI voice calls across 13+ organizations, achieving 66% performance improvement and 40ms webhook latency.
- Engineered comprehensive analytics system with 11 RESTful API endpoints and automated credit management, whilst optimizing PostgreSQL database through 43 strategic indexes—resulting in 90-95% query time reduction and implementing JWT-based security, Chargebee integration, and Azure Blob archival.

AI-ML Intern, Bigship Technologies February – April 2025

- Architected and integrated a comprehensive AI-centric voice infrastructure, demonstrating expertise in real-time STT, advanced LLM processing, and TTS synthesis to establish a foundation for automated, high-fidelity voice interactions.
- Led the design and successful deployment of a production-grade proof-of-concept for intelligent voice response, expertly fine-tuning and integrating LLaMA models to address complex, company-specific use cases.

Tech Intern, Corporate Infotech Pvt Ltd (CIPL) July – August 2024

- Innovated by fine-tuning a custom GPT-2 LLM on 350+ HR queries and developing an intuitive Flask-based UI, resulting in a 40% reduction in average HR query response time, directly improving employee support and satisfaction.

Skills

Programming AI/ML: Python, LLMs (GPT-4, LLaMA, Gemini), Generative AI, NLP, PyTorch, TensorFlow, Transformers, RAG, AI Agents, Prompt Engineering, MLOps
Backend Infrastructure: Vector Databases(Pinecone), PostgreSQL, REST APIs, Microservices, JWT, Webhooks,
Voice Technologies: LiveKit, Telephony Integration (Twilio, Plivo), Voice Avatars, Real-Time Conversational AI, Real-time STT/TTS, DTMF, Voicemail Detection, VAD, Turn Detector

Cloud DevOps: AWS, Azure, Git, CI/CD, Docker

Projects

InsightGen: Multi-Agent Market Research and Use Case System InsightGen

- Configured Multi-Agent System leveraging Generative AI (Gemini) and Serper.dev to automate market research, AI/ML use case analysis, and resource collection, whilst engineering feasibility and ROI evaluations to enhance operational efficiency.

SnapDetect AI: Image Analysis System SnapDetect AI

- Built an AI-powered image analysis system integrating YOLOv5x, Tesseract OCR, and Mask R-CNN to achieve 95% object detection accuracy.
- Implemented BART NLP for summarizing image attributes, enabling efficient object identification through a user-friendly web interface.

LLM-Powered HR Chatbot LLM-Powered HR Chatbot

- Created a full-stack, AI-driven HR chatbot with a fine-tuned GPT-2 model, Flask backend, and a responsive React frontend with Framer Motion for a modern user experience.

Melanoma Detection Using Deep Learning (CNN Model) Melanoma Detection CNN

- Designed and trained a CNN model in TensorFlow/Keras to accurately detect and classify melanoma from skin images, achieving up to 90% diagnostic accuracy.

Certifications

Oracle Cloud Infrastructure 2025 Certified AI Foundations Associate Sep 2025
Certification - "I built a virtual agent at IBM TechXchange Dev Day" – IBM Jan 2025
Building Your first RAG System using LlamaIndex – Analytics Vidhya Nov 2024
Deep Learning With PyTorch – IBM Nov 2024
IBM Z Day 2024 - AI & Data Certificate – IBM Oct 2024
Python And Django Framework For Beginners Complete Course – Udemy Jul 2023

Education

B.Tech CSE (AI-ML), Uttaranchal University 2021 – 2025
Class 12th, H.R. Public School 2019 – 2020