

Suyash Bhavsar

AI Engineer | Voice AI & Real-Time Conversational Systems | LLM & RAG | Backend Systems

Pune, India | bhavsarsuyash30@gmail.com | +91 7350916215 | LinkedIn | GitHub

Skills

Voice AI

Real-Time Voice AI (OpenAI Realtime API, Gemini Live API) • Full-Duplex Audio Streaming • WebRTC/WebSocket/SIP Audio Transport • Live-Agent Hand-off • Call-Quality & Latency Monitoring

AI / LLM

Large Language Models (LLM) • Retrieval-Augmented Generation (RAG) • Hybrid Search (structured filtering + semantic/vector search) • Model Context Protocol (MCP) • LangChain • Prompt Engineering • Conversational AI • Vector Databases (Pinecone) • LLM API Providers (OpenAI, Google Gemini) • LLM Evaluation (manual review, A/B testing, production quality monitoring)

Backend

Async Programming (asyncio) • Python (FastAPI, Flask) • Spring Boot • REST APIs • Microservices • Multi-Tenant Architecture • SQL

Systems & Real-Time

Event-Driven Systems • Redis (Celery broker, caching, distributed locking) • Celery • Workflow Orchestration • WebRTC • WebSockets

Observability & Monitoring

OpenSearch-based logging & analytics • Real-Time Alerting (Google Chat)

Cloud & DevOps

AWS (CodeDeploy) • Docker

Data

PostgreSQL • MySQL • MongoDB • OpenSearch

Tools

Git • GitHub • Linux • Postman • Jira • Claude Code (AI coding agent)

Education

MIT World Peace University

M.Sc. Computer Science

2022 – 2024 | GPA: 8.06

Savitribai Phule Pune University (SPPU)

B.Sc. Computer Science

2019 – 2022 | GPA: 8.33

Work Experience

AI Engineer

Pune, India

Graas

Jan 2024 – Present

- **Singlehandedly** designed, built, and deployed Graas's **full-duplex, low-latency** real-time voice AI system **end-to-end** — using **WebRTC/WebSocket** with a pluggable AI provider layer (**OpenAI Realtime, Gemini Live**) — across WhatsApp and Facebook, driving **70% of interactions** through messaging channels, with seamless live-agent hand-off mid-call
- Scaled the voice system to power outbound calling campaigns reaching **5,000+ customers**, backed by a live call-quality and latency monitoring system that brings root-cause debugging of production voice issues down to **under 14 minutes**
- Developed and deployed an LLM-powered conversational AI system supporting general chat-based workflows across use cases (not limited to e-commerce), handling **1000+ daily user queries**, including product discovery, cart, and order automation as one deployment
- Designed and implemented a **hybrid product-retrieval system** for conversational search — combining structured query parsing (DSL-based filtering), **KNN semantic search**, and a multi-tier fallback chain (structured → vector → keyword) — improving retrieval relevance and reducing irrelevant/fallback responses
- Built scalable backend services in **Python (FastAPI)** using an async ASGI architecture and gevent-pooled Celery workers, supporting **500+ concurrent conversational sessions**; reduced response latency via **Redis-based asynchronous processing** — decoupling webhook ingestion from LLM response generation through a durable, queue-based pipeline, with per-organization LLM prompt caching and short-TTL product caching to cut redundant OpenSearch/embedding calls
- Engineered a versioned **hook-based customization framework** paired with a UI-based function runner, used by **10+ client organizations** to create custom tools, configure scheduled jobs, and apply business-logic changes — all without modifying platform code, redeploying, or engineering involvement

Projects

AICompete | AI Competitive Intelligence Platform (GitHub)

- Built an end-to-end AI competitive intelligence platform using **FastAPI, Celery**, and scraping pipelines, implementing LLM-powered summarization and a **RAG pipeline with Pinecone embeddings** for contextual knowledge retrieval
- Developed background analytics and alerting workflows using **Scrapy, Redis, and PostgreSQL**, enabling real-time sentiment tracking, keyword detection, and automated competitor monitoring

Achievements

- **Publication** — “Predicting Student Placement using ML Models,” IJFMR (May 2024); Random Forest achieved **81.47%** accuracy
- Qualified **UGC-NET & Maharashtra SET** (June 2024) for Assistant Professor Eligibility, both on the **first attempt**
- Solved **150+ problems** on LeetCode, reinforcing data structures and algorithms fundamentals