

TARTENDU KUMAR

Lucknow, Uttar Pradesh

+91-7589454833

AI Systems Engineer · Voice AI Systems · Enterprise AI Infrastructure

tartendukumar@gmail.com

zencia.ai · tartendu.in

SUMMARY

AI Systems Engineer with **2+ years** building and shipping production AI infrastructure — solely, end-to-end. Founding engineer at **Zencia.ai**: designed and delivered a live multi-tenant SaaS platform, a production voice AI agent (**echo.zencia.ai**, 40+ languages), and a GCP-deployed AI demo agent platform (**envoy.zencia.in**) — all built alone from architecture to deployment. Core expertise: real-time voice agents, RAG pipelines, offline LLM systems, and multi-tenant AI product engineering. Specialisation in **privacy-first, on-premise AI** — zero cloud dependencies.

TECHNICAL SKILLS

Languages	Python, JavaScript, TypeScript, SQL, HTML5, CSS3
AI / ML	LLMs, RAG Pipelines, FAISS Vector DB, SentenceTransformers, Embeddings, Ollama (local inference), LangChain, MCP (Model Context Protocol), NLP, Prompt Engineering
Voice AI	ElevenLabs Conversational AI, Twilio (Voice / SMS / WhatsApp), Real-time PCM Audio Streaming, Barge-in Handling, Post-call Webhook Processing, Local TTS (HuggingFace Transformers · PyTorch · diffusion-based streaming inference · CUDA / MPS)
Backend	FastAPI, Flask (Blueprints), asyncio, WebSocket, REST APIs, RBAC, Rate Limiting, uvicorn
Frontend	React.js, TypeScript, Vite, Responsive Web Design
Cloud & DevOps	GCP (VM deployment), AWS Elastic Beanstalk, Nginx (reverse proxy + SSL), systemd, Linux, Let's Encrypt, Git
Databases	Firebase / Firestore, SQLite, FAISS
AI APIs	Gemini API, OpenAI API, ElevenLabs API, Groq API, Together AI
Integrations	Microsoft Graph API (MSAL · Outlook Calendar), sell.do CRM API, Google Sheets API, Google Calendar API, Google Meet Bot, Chrome Extension, Firebase Admin SDK, Twilio SDK, pywebview / PyInstaller

PROFESSIONAL EXPERIENCE

AI Systems Engineer Oct 2024 – Present (1 yr 7 mo)

Zencia.ai · Lucknow, India Sole Developer — Frontend to Backend

- Zencia SaaS Platform** (zencia.ai) — sole architect and developer of a live multi-tenant AI platform.
 - ›17-module Flask Blueprint codebase: agent management, knowledge base, call logs, billing, phone numbers, SMS, calendar, Google Sheets, client portal, partner API, admin panel
 - ›Billing system with plan tiers, balance enforcement, low-balance alerts, and embeddable JS widget for client sites
- Zencia Echo** (echo.zencia.ai) — production voice AI agent supporting **40+ languages**.
 - ›Conversational AI + persistent customer memory injected into agent context per call
 - ›Post-call webhook handler for transcription storage and deduplication; SMS/WhatsApp follow-up via Twilio
- AI Demo Agent Platform** (envoy.zencia.in) — interactive product demo tool deployed solo on GCP.
 - ›Conversational AI voice guide walks users through recorded product demos; AI avatar renders lip-sync via audio passthrough
 - ›Per-demo voice agent agent with dynamic system prompts; AI navigates slides via go_to_slide tool calls in real time
 - ›FastAPI + React/TypeScript/Vite; GCP VM with Nginx reverse proxy, SSL (Let's Encrypt), systemd service
- Voice Agent** — standalone real-time AI voice agent built with API.
 - ›FastAPI + asyncio WebSocket server; bidirectional PCM audio streaming (16 kHz in / 24 kHz out)
 - ›Async session registry: **50 concurrent sessions**, idle reaper, auto-reconnect with exponential backoff, barge-in/interruption support

- **Enterprise RAG System** (Tender AI) — end-to-end document intelligence, built alone.
 - › FAISS vector DB + SentenceTransformers (all-MiniLM-L6-v2) embeddings + Ollama local inference (llama3, mistral, phi3)
 - › PDF/DOCX/TXT ingestion pipeline with chunking, RBAC (3 tiers), and background auto-reindex thread
- **Zencia Edge** — offline enterprise AI desktop app; zero cloud dependencies.
 - › Packaged as Windows executable via pywebview + PyInstaller; LangChain reasoning + MCP tool integrations
- **Hyundai AI Voice Platform** — complete system designed and presented to Hyundai.
 - › Inbound/outbound AI calling, lead scoring (Cold/Warm/Hot), test-drive booking, geo-intelligent agent routing
 - › CRM integration, call recording + AI transcript + SMS follow-up, real-time analytics dashboard
- Built **MCP (Model Context Protocol)** tools: patient search, drug interaction checker, email automation agent, document chatbot.
- **Zencia Vocalis** — local on-premise voice pipeline; zero cloud TTS dependency.
 - › Integrated Microsoft VibeVoice (open-source streaming TTS, ~300 ms first-chunk latency) as a local TTS engine — on-device alternative to ElevenLabs
 - › Deployed and tested across CUDA / CPU / MPS hardware; streamed text input for near real-time speech generation

Tech Executive

Dec 2023 – Jun 2024 (6 mo)

Zotation Pvt Ltd · Lucknow, India

- Developed and maintained 5+ client websites (Python backend, HTML/CSS/JS frontend).
- Built business automation workflows reducing manual processing time for clients.
- Assisted in AI-powered customer service solution development.

KEY PROJECTS

Zencia AI SaaS Platform

LIVE — [zencia.ai](#)

Solo Built

Python · Flask Blueprints · Firebase/Firestore · ElevenLabs · Twilio · Google APIs · React.js · AWS

- Multi-tenant SaaS: owner accounts, client portals, admin panel with revenue and analytics dashboards.
- AI voice agent creation ; per-agent knowledge base with document ingestion and retrieval.
- Billing enforcement, plan limits, partner API, and JS embed widget — all production-deployed.

AI Demo Agent Platform

LIVE — [envoy.zencia.in \(GCP\)](#)

Solo Built

FastAPI · React/TypeScript/Vite · ElevenLabs Conversational AI · Firebase/Firestore · GCP VM · Nginx · SSL

- Interactive product demo platform: users record screenshot walkthroughs; AI voice agent guides visitors through them conversationally.
- AI avatar renders real-time lip-sync via audio passthrough from voice agent (pcm_16000); AI navigates slides via `go_to_slide` tool calls.
- Per-demo voice agent created dynamically with system prompts built from demo steps; `agent_id` cached in Firestore.
- Deployed: GCP VM → Nginx (SSL, Let's Encrypt) → uvicorn → FastAPI → React/Vite frontend; managed via systemd.

Live Real-time Voice Agent

Solo Built

FastAPI · asyncio · WebSocket · Live API · Python

- Standalone real-time voice agent — bidirectional PCM audio streaming (16 kHz in / 24 kHz out) via Live API over WebSocket. Async session registry capped at **50 concurrent sessions**; idle reaper (5-min timeout), auto-reconnect with exponential backoff,
- barge-in/interruption handling.

Enterprise RAG Document Assistant

Solo Built

Python · Flask · FAISS · SentenceTransformers · Ollama · PyMuPDF · python-docx

- Full RAG pipeline: ingest → extract → chunk → embed (SentenceTransformers) → FAISS index → retrieve → generate (Ollama local LLM).
- Multi-model at runtime (llama3 / mistral / phi3 / neural-chat); RBAC with 3 permission tiers; background auto-reindex.

Zencia Vocalis — Local On-Premise Voice Pipeline Solo Built

Python · Microsoft VibeVoice · HuggingFace Transformers · PyTorch · CUDA / MPS · Streaming TTS

- Integrated Microsoft's open-source VibeVoice streaming TTS model as a zero-cloud, on-device alternative to ElevenLabs — no audio data leaves the device.
- ~300 ms first-chunk latency with streaming text input; tested on CUDA, CPU, and MPS hardware.
- Built as the TTS engine for Zencia Vocalis — a fully local voice AI pipeline with no cloud dependencies.

Zencia Edge — Offline AI Desktop App Solo Built

Python · Flask · LangChain · MCP · pywebview · PyInstaller · Windows Edge WebView2

- Packaged as a Windows executable; Flask backend + pywebview UI — no browser install needed.
- LangChain reasoning pipeline + MCP tool integrations; fully offline, zero cloud dependencies.

EDUCATION

Bachelor of Computer Applications (BCA) 2023 · 83%
BabuBanarasiDasUniversity · Lucknow,India
Data Structures, Algorithms, Database Management, Software Engineering, Computer Networks

High School & Intermediate (CBSE) 2020 · 85%
KendriyaVidyalayaLakhanpur · Jammu,India

ADDITIONAL

Languages: English (Fluent), Hindi (Native) **GitHub:** Private repositories (available on request) **Methods:** Agile, API-first, modular architecture, Git