

HEMENDRA TRIPATHI

AI Engineer · Multi-LLM & Voice AI Systems · Technical Lead

Udaipur, Rajasthan, India | 063787 45958 | hemendratripathi880@gmail.com

[Portfolio](#) | [LinkedIn](#) | [GitHub](#)

SUMMARY

AI Engineer and Technical Lead with 4+ years building and shipping production AI systems — multi-LLM orchestration, RAG pipelines, and real-time voice agents. Currently leading engineering for **Callin.io**, an AI voice platform serving 1,500+ paying customers, where I designed a **complexity-aware model-routing layer** that cut LLM inference cost by 20% while cutting time-to-first-token to ~380ms. Hands-on across the full AI product lifecycle: retrieval (Pinecone, Supabase Vector), voice cloning and real-time telephony, usage-based billing for AI products, and leading a 5-engineer team from architecture to revenue.

AI ENGINEERING SKILLS

LLM Orchestration: Multi-LLM routing by task complexity & cost, semantic caching, concurrent prompt execution, prompt engineering, TTFT optimization

RAG & Retrieval: RAG pipeline design, Pinecone, Supabase Vector, embeddings, context/window management

Voice & Real-Time AI: Real-time voice agents, voice cloning (ElevenLabs, Cartesia), streaming audio over WebSockets, telephony integration (Twilio, Telnyx, SIP)

AI Product Engineering: Usage-based billing for AI features, Stripe metering, cost/latency tradeoff design, infra spend ownership

Full-Stack & Platform: React, Next.js, React Native, Node.js, NestJS, Express, Python, Laravel

Cloud & Data: AWS, Docker, PostgreSQL (Supabase), Redis, CI/CD, event-driven microservices

Integrations: Google OAuth/Calendar, SendGrid, Firebase/FCM, multi-provider email sync, n8n, Zapier

Leadership: AI architecture decisions, 5-engineer team leadership, hiring, client acquisition, mentorship (150+ students)

EXPERIENCE

Technical Lead / AI Engineer Oct 2024 – Present

Appspundit Infotech — Udaipur, Rajasthan | Client engagement: Callin.io, an AI voice-calling SaaS platform

- Architected the **multi-LLM orchestration layer** powering all AI products — routing each conversational turn to the smallest capable model, cutting inference cost 20% while reserving larger models for complex reasoning.
- Built a **semantic caching and concurrent-prompt execution** system that reduced voice-agent time-to-first-token to ~380ms, keeping real-time calls natural under load.
- Designed **RAG pipelines** (Pinecone, Supabase Vector) and real-time telephony integrations (Twilio, Telnyx, SIP) for multilingual voice agents with voice cloning and contextual prompting.
- Scaled the platform to **1,500+ paying customers**, including enterprise accounts in healthcare and real estate, integrating usage-based billing (Stripe) with the AI product surface.
- Extended the multi-LLM architecture across two additional products — **CondoMail** (AI email triage/drafting) and **Realead** (AI lead-qualification calling) — as Technical Lead over a 5-engineer team, reporting directly to the founder.
- Owned customer acquisition and retention, converting 30+ prospects into long-term paying accounts across healthcare and real estate.

Freelance Full-Stack Developer 2022 – 2024

Remote — fintech, retail, logistics clients

- Delivered 10+ end-to-end applications across MERN, Django, Flask, and Laravel, including AI-adjacent automation and API-driven systems.
- Built a FinTech accounts management system for an MCA-registered firm, automating reconciliation and saving 15+ staff-hours per week.
- Improved API performance by 25% and implemented CI/CD pipelines for zero-downtime deployments.

Technical Instructor 2021 – 2023

Aimers Institute & VT College — Udaipur / Bhilwara, Rajasthan

- Mentored 150+ students in Python, Django, MERN, and Flutter through project-based learning.
- Redesigned course curriculum aligned to industry needs, contributing to a 42% increase in student placement rates.
- Supervised 30+ capstone projects, instilling best practices in version control, API design, and UI development.

SELECTED AI PROJECTS

[Callin.io](#) — AI Voice-Calling Platform (*React, Node.js, Supabase, Twilio, Telnyx, Stripe, Firebase*)

- Engineered a low-latency telephony pipeline using WebSockets and SIP for real-time audio streaming; built the multi-LLM orchestration layer with semantic caching and concurrent prompt execution to minimize TTFT for voice agents.

[CondoMail](#) — AI Email Management (*React, Node.js, Supabase, Stripe, Firebase*)

- AI agents that classify, draft, and auto-reply to email across multiple providers; live in production with early adopters including high-volume Amazon sellers.

Realead — AI Real Estate Lead Qualification (*React Native, NestJS, Supabase, Stripe*)

- AI-driven calling agent that qualifies and follows up with new property leads from connected lead sources; in final beta ahead of release.

Sunria & FinTech Internal Accounts (*Laravel, Flutter, MERN*)

- Pan-India farm management system with real-time field-to-warehouse communication, and a secure financial dashboard with automated reconciliation that reduced accounting errors by 30%.

EDUCATION

Master of Computer Applications (MCA) — Janardan Rai Nagar Rajasthan Vidyapeeth (*Pursued part-time alongside full-time engineering leadership roles; Expected 2026*)

Bachelor of Computer Applications (BCA) — Mohanlal Sukhadia University, Udaipur (2022)

LANGUAGES

English (Fluent), Hindi (Native)