

MADHAV VISANI

AI & Full-Stack Engineer

✉ madhavvisani10@gmail.com | 🌐 [See the Work](#) | 📄 [Madhav Visani](#) | 🎮 [GeneralTony007](#)

Gandhinagar, India · **Open to remote roles worldwide** (US / EU overlap) · Open to relocation

SUMMARY

AI and full-stack engineer who builds LLM products end to end — retrieval, reliability, evaluation, and adversarial testing — and ships them to production. Shipped **production AI assistants for paying clients**, grounded in each client's own website data and hardened with fallback-model routing and rate limiting, plus an end-to-end internal tool for a manufacturing client. Independently designed and deployed five systems, including an autonomous LLM red-teaming platform implementing **12 published attack papers** and a four-client cross-device system running as daily personal infrastructure.

EXPERIENCE

Full-Stack Developer & AI Integration Intern

Oct 2025 – Dec 2025

Volterx Infotech

On-site — Ahmedabad, India

- Built and shipped **multiple production AI assistants** for client websites, combining OpenAI API integration with a from-scratch retrieval layer over each client's own site content, so answers came from the business's real data instead of a generic model.
- Engineered the serving path for reliability and cost control: **automatic fallback to a secondary model** when the primary provider failed or throttled, per-user rate limiting, and typed service boundaries that kept prompt and provider changes isolated from application logic.
- Helped build an **end-to-end internal tool for a manufacturing client** — React/TypeScript frontend, backend APIs, and data layer on the live production codebase alongside senior engineers — carrying it from requirements through delivery, onboarding, and triage of issues raised in daily commercial use.

SELECTED PROJECTS

WRAITH — Autonomous AI Red-Teaming Platform

wraith-redteam.vercel.app

Python 3.12, FastAPI, Next.js 16, PostgreSQL, Redis, Docker

- Built a platform that generates and runs adversarial attacks against LLM applications: **12 attack strategies** spanning black-box jailbreaks (PAIR, TAP, AutoDAN, Crescendo, Best-of-N) and injection simulators (indirect/RAG, tool-hijack, PDF, vision), each implemented from the paper that introduced it.
- Unified **8 model providers** (OpenAI, Anthropic, Gemini, Together, HuggingFace, vLLM, Ollama) behind one interface, and made every finding reproducible and tamper-evident: seed, target model and attack version **Ed25519-signed**, with harm classification mapped to the OWASP LLM Top-10 and CWE.
- Kept the Python engine torch-free and CLI-runnable so the deployable image fits a free tier; **121 tests across 19 modules**, CI on every push, attacks streamed live to the dashboard over SSE.

Universal Link Vault — Cross-Device Capture System

closed-source — holds live personal data

Next.js 16, React 19, TypeScript, Supabase (Postgres/Auth/RLS), Upstash Redis, Chrome MV3, Win32 Python

- Designed, built and now run as **daily personal infrastructure** a cross-device capture system spanning **four separate clients** — a native Win32 tray watcher, an installable Android PWA shipped as a signed package, a Chrome MV3 extension, and the web app — writing into one Supabase store, searchable from a fuzzy command palette.
- Enforced a private section **at the database rather than in the UI** — those rows stay unreadable and unwritable even while holding a live app session token, and access requires a second-factor PIN minting a short-lived, login-bound token held only in memory — alongside passkey auth, a session cap, and a remote kill switch.
- Added a cron-scheduled ingestion pipeline that pulls and normalizes opportunities from **20+ sources** (Upwork, Freelancer, YC Work at a Startup, Greenhouse/Lever/Ashby boards, Hacker News) into one deduplicated, classified feed with push notifications and a scheduled digest.

Market Pulse — AI Review Intelligence

market-pulse-two-phi.vercel.app

Next.js 16, TypeScript, Supabase (Postgres + pgvector), Gemini Flash, Zod, Vitest

- Built an AI review-intelligence product for e-commerce brands: it ingests public product reviews, embeds them into pgvector, then clusters them into named themes and ranks the top fixes customers are asking for — resolving **4,818 real reviews into 25 themes** across 3 brands, with a weekly competitor price watch and an automated digest.

- Wrote an **offline evaluation harness** scoring the sentiment model against a human-labeled golden set (accuracy, macro-F1, confusion matrix) and exposed it as a live endpoint — behind a resumable, idempotent ingestion state machine with per-call quota rate-gating and PII hashing.

Nexus Quant — Browser-Based Market Terminal

nexus-quant-ochre.vercel.app

Next.js 15, React 19, TypeScript, D3.js, MapLibre GL + deck.gl, Firebase

- Built a Bloomberg-style market workstation in the browser: a command-driven, multi-panel terminal with **20+ mnemonic functions** over live data from Yahoo Finance, Binance, Finnhub and NSE, plus a Gemini copilot and a real-time vessel-radar map over live AIS traffic.
- Enforced a strict data-provenance model badging every value LIVE / DELAYED / MODELED / SIM / DEMO so nothing synthetic can read as real, and shipped a paper-trading league with a margin engine and RBAC admin tooling.

Lead Rescue — AI Missed-Lead Recovery

leadrescue-hq.vercel.app

Next.js 16, TypeScript, Supabase (Postgres + pgvector), Gemini Flash, Resend, Vercel

- Deployed a niche-agnostic AI lead-recovery widget for service businesses: a RAG chatbot answers visitors from the business's own documents, qualifies leads and books appointments, and degrades to a plain capture form when the LLM is unavailable so a lead is never silently lost. *(Live link runs a demo tenant.)*

Every project above can be walked through live on a screen-share — code, commit history, and architecture — at any point in an interview.

TECHNICAL SKILLS

Languages: Python, TypeScript, JavaScript, SQL, Go

AI & LLM: LLM application engineering (OpenAI, Anthropic, Gemini), RAG and vector search (pgvector), structured/JSON-mode prompting, provider failover and rate limiting, LLM evaluation (golden sets, macro-F1, confusion matrices), adversarial red-teaming (PAIR, TAP, AutoDAN), self-hosted inference (Ollama, vLLM)

Backend: FastAPI, Pydantic v2, Node.js, SQLAlchemy 2, Alembic, REST APIs, Server-Sent Events

Frontend: Next.js, React, Tailwind CSS, D3.js

Data & Infrastructure: PostgreSQL, Redis, Supabase, pgvector, Docker, Vercel, GitHub Actions (CI)

Security & Testing: OWASP LLM Top-10 and CWE mapping, row-level security, passkey/JWT auth, Ed25519 signing, CSP/HSTS, pytest, Vitest, Git

EDUCATION

LDRP Institute of Technology & Research

Gandhinagar, India

B.Tech, Computer Engineering | CGPA: 7.89/10

Graduated April 2026

CERTIFICATIONS & PROGRAMS

- **Python for Data Science** — NPTEL, IIT Madras
- **Google Cloud Arcade — Champion** — Google Cloud Skills Boost, across a season of hands-on Google Cloud labs. [Verified profile](#)
- **Pre-Release Beta Tester — global smartphone manufacturer (program under NDA):** selected into a manufacturer partner program to run pre-release Android builds on unreleased flagship hardware, filing defect reports, stability regressions, and UX feedback ahead of public launch. *(Manufacturer, program name, and device models withheld under NDA.)*