

HARSHAVARDHAN JOTHIKUMAR

Founding-style engineer: real-time voice AI, LLM evals and retrieval, shipped end to end

+91-7358975075 | harshavardhanjo@gmail.com | harshavardhanjo.com | GitHub | LinkedIn | Bangalore

SUMMARY

Engineer who takes products from an empty repo to paying users. At **Mynaksh** I built a real-time AI voice product to a paid live service in ~6 weeks (five services, 11 languages, interruption handling at sentence granularity) and **Shared Brain**, a per-user temporal knowledge graph in Neo4j that is now the long-term memory behind ten product surfaces, with recall lifted from 65% to 99% by measurement. Previously **Founding Engineer at Mantys (YC W23)**: LLM evaluation to 98% accuracy and \$2.3M in claims recovered. 3+ years shipping production systems, twice as a founding engineer.

EXPERIENCE

SDE II (Fullstack, AI)

Jan 2026 – Present

Mynaksh

Bangalore, India

- **AI Voice Product — Empty Repo to Live Paid Service in ~6 Weeks:** Built the company's real-time AI voice product end to end across five services — holding a 20ms audio cadence against second-scale LLM latency with per-minute wallet billing, interruption handling that tracks what the caller actually heard at sentence granularity, streaming speech recognition and synthesis across 11 languages, and per-turn model calls cut from four to one by making the conversational model the orchestrator.
- **Shared Brain — Per-User Temporal Knowledge Graph (Neo4j), 38 modules / ~12k lines:** Designed and shipped the long-term memory system now serving ten product surfaces including live voice, replacing three divergent flat memory formats with one retrieved block. World time is modelled separately from system time, with four distinct supersession kinds so a corrected fact is never resurfaced while a refined one is subsumed rather than falsified.
- **Retrieval and Evaluation — Recall 65% → 99%:** Built recall as a union of vector cosine, BM25 keyword and entity branches with no model call in the path, and let measurement overturn the original design: scoping the vector branch to matched episodes capped recall at **65.4%**; removing the gate lifted it to **98.9%**. Fitted fusion weights across 192 sets with two-fold holdout — plain cosine beat hand-tuned weights and Reciprocal Rank Fusion.
- **Four-Stage Evaluation Harness:** A labeller that marks the relevant set **blind to what retrieval returned**, and contradiction treated as a **hard gate rather than a score tradeable against recall** — because a missed fact is invisible and a wrong one is a trust event.
- **Owned the Paid Call System and Its Money Path:** Introduced calls as a first-class entity and built the vendor web-hook handlers that reconcile a third-party event stream into a billable session — idempotent wallet settlement, call-end recovery when the vendor SDK failed outright mid-transaction, and the presence and busy-state logic governing marketplace supply. Measured fully-loaded cost per conversation to produce the unit-economics model behind pricing.
- **Root-Caused Every Call Dying at Exactly Ten Minutes:** Recorded in telemetry as ordinary user hangups, so the system looked healthy — traced to a global-instead-of-per-channel refresh timer in the WebRTC library's TURN layer, confirmed by enabling library-internal tracing in production, and fixed with a vendored patch.
- **Stack:** Python (asyncio, FastAPI, PyAV/Opus), Gemini structured output & context caching, OpenAI embeddings, Neo4j/Cypher, Node.js/TypeScript microservices, Redis, PostgreSQL, BullMQ/SQS, WebRTC, AWS ECS Fargate, Grafana/Loki.

Founding Engineer

Mar 2025 – Sep 2025

Mantys Healthcare AI • Y Combinator W23

Bangalore, India

- **LLM Evaluation Framework:** Built the evaluation system validating LLM extraction accuracy for copay and eligibility fields — golden sets, judged scoring and regression checks — cutting manual verification by 80% and lifting overall system reliability to **98% accuracy**.
- **Prior Authorization AI:** Built a prior-authorization decision system from scratch automating workflows across **120k+ medical procedures**, replacing specialist teams of 8–10 with 2–3 staff.
- **Claims Revenue Recovery:** Architected the financial reconciliation system that recovered **\$2.3M** in pending claims, automated status tracking for 50k+ monthly claims, and cut reconciliation errors by 90%.
- **Technologies:** Python, FastAPI, Celery, Playwright, PostgreSQL, Docker, AWS, LLM API integration.

Founding Team Member & Chief Technology Officer

Nov 2023 – Mar 2025

Reachgig

Chennai, India

- **Full-Stack Platform:** Built and launched a freelancing marketplace from scratch — Next.js 14, Node.js, TypeScript and AWS serverless (Lambda, EC2, DynamoDB, S3) with payments, auth and real-time messaging — 99.8% uptime with auto-scaling that cut infrastructure cost by 40%.
- **Mobile & Delivery:** Shipped the cross-platform React Native/Expo app, and built the CI/CD pipeline cutting deployment time by 70% alongside a notification system at a 99.5% delivery rate.

Product Engineer

Sep 2022 – Mar 2023

Kuddle Pet Care

Bangalore, India

- Led Next.js 13 development against a Kotlin microservice backend — caching and optimisation improved page load times 55% and lifted engagement 40%. Architected a social platform from scratch (Flutter + Kotlin), presented it to the executive board and shipped it — 35% retention increase, 25% DAU boost.

TECHNICAL SKILLS

Languages: Advanced: Python, JavaScript/TypeScript • Proficient: Go, Java, C++ , Kotlin

Applied LLM: Agent orchestration • Structured output under schema constraints • Context caching • Model routing by task, cost & latency • LLM evaluation & LLM-as-judge • Grounding & hallucination control

Retrieval & Memory: Neo4j/Cypher • Temporal knowledge graphs • Vector search & embeddings • Hybrid BM25 + semantic retrieval • Rank fusion • pgvector/RAG

Backend & Data: FastAPI • asyncio • Node.js • Celery • BullMQ/SQS • PostgreSQL • Redis • MongoDB • Microservices • Idempotent settlement • Wallet/ledger design • Retry, dead-letter & replay

Real-Time & Voice: WebRTC • SRTP/DTLS/ICE • STUN/TURN • Opus/RTP • Jitter buffering • VAD & barge-in • Streaming STT/TTS

Cloud & Observability: AWS (ECS Fargate, Lambda, EC2, S3, DynamoDB) • GCP • Docker • CI/CD • Grafana/Loki • Prometheus • On-call & incident response

EDUCATION

B.Tech, Computer Science & Engineering

2019 – 2023

Vellore Institute of Technology, Chennai

CGPA: 8.45/10.0

- Top 6% of 2,500 students; selected for a pre-placement offer. Tech Head, Data Structures & Algorithms club (200+ members).
- **Certifications:** AWS Cloud Practitioner (CLF-C01), 2023.