Govind Sharma

Senior Full Stack AI Engineer · LLM Apps, React & Node.js

govindgovind852@gmail.com · Mumbai, India · govindio.com · linkedin.com/in/govind-io · github.com/govind-io

Open to Senior or Lead Full Stack and AI Engineer roles · Remote (US/EU overlap) or India

SUMMARY

Senior full stack AI engineer with 5+ years shipping LLM-powered products end to end. At Stamina.io, built **Zara**, an AI SDR agent that enriches **20K leads/day** and doubled reply rates, and the workflow engine behind **5,000+ workflows/day**, while ARR grew from **$250K to $2M+** in 9 months. Built **FastBowlingLab** solo, an AI coaching app on web, iOS and Android with **1,200+ real users**. Before that, LLM-powered retrieval over 1M+ documents for Fortune 500 sales teams and a React SDK and WebRTC platform for 100,000 concurrent users. Works across **React/Next.js**, **Node.js/NestJS**, **Python** and **LangGraph**.

EXPERIENCE

Staff Engineer · Full-time Contract | Stamina.io | Jun 2025 – Present

Promoted: Founding Backend Engineer — Lead (Jun 2025) → Staff Engineer (Apr 2026)

AI sales, CRM and marketing SaaS · **1,000+ customers** · **50+ countries** · **$2M+ ARR** · New York, United States · Remote

› Designed and built **Zara**, Stamina's AI SDR agent (**LangGraph**, tool calling, structured outputs): enriches **20K leads/day**, doubled reply rates and cut lead response time from 24 hours to instant for 1,000+ customers in 50+ countries.

› Cut LLM cost per lead by **70%** with per-task model routing: bulk per-row work on small models, larger models only for classification and spec generation, with **LLM evals** and **guardrails** (schema-validated outputs, model fallback, confidence-gated human escalation) holding output quality.

› Architected the workflow automation engine (**5,000+ workflows/day**, 50+ triggers and actions, 10+ integrations) that starts exactly one run per event on at-least-once delivery: transactional outbox to **Google Pub/Sub**, layered dedup in **PostgreSQL** and a durable step machine with leases and version pinning.

› Designed the credit ledger behind usage billing: serializable, idempotent spends across concurrent workers, expiring grants spent soonest-first, an append-only ledger enforced in the database, and **Stripe** sync that survives late and duplicate webhooks.

› Built the **Twilio** power dialer used by **200+ customers making 5,000+ calls/day**: row-locked lead claims so no lead is dialled twice, signed per-call tokens re-checked in the voice webhook, and per-brand Twilio subaccounts.

› Shipped **SSO/OAuth2**, **RBAC** and an OAuth 2.1 **MCP server** for larger accounts, on **Kubernetes**, **Terraform** and AWS with per-PR preview environments and **99.9%+ uptime**.

› Promoted from founding engineer to Staff Engineer in 10 months; led the **10-engineer team** through design reviews, weekly releases and on-call.

Software Engineer II — Engineering Lead | Better Software · Embedded at NeshAI | Jul 2024 – Jun 2025

B2B AI SaaS for Fortune 500 sales teams · Remote

› Engineering lead for a **5–6 engineer** team; built NeshAI's AI product recommendation engine: **LLM-powered retrieval** over 1M+ documents (**pgvector** + **Elasticsearch**), LinkedIn enrichment and ICP analysis for **Blue Origin, Chevrolet, DKSH, WRGrace**.

› Held p95 retrieval under **800ms** at 85% answer accuracy on the eval set while ingesting 1,000+ new docs/day through event-driven **Kafka** pipelines.

› Shipped **MCP servers**, OpenAI Assistants API integrations and AI workbooks rated **5 stars** by sales teams at $1B+ revenue enterprises.

› Built secure connectors for **Salesforce, Google Drive, Amazon S3 and Azure Blob Storage**, and implemented the **SOC 2** controls enterprise clients required.

Sr. Software Engineer — Tech Lead | Loktantra Mediatech | Jan 2022 – Jun 2024

Promoted twice: Software Engineer (Jan 2022) → Team Lead (Jul 2022) → Sr. Software Engineer — Tech Lead (Jan 2023)
Social and live-streaming platform (Khulke) · Bengaluru, India (Hybrid)

› Built a self-hosted WebRTC SFU on **LiveKit**, a signaling server and a React SDK from scratch, replacing a **$2,000/month** vendor; scaled to 100,000 concurrent users across the platform at 1,000 req/sec, with rooms of 50+ speakers and 500+ listeners.

› Cut page load from **4.2s to under 1s** and API p95 from 900ms to 200ms through Redis caching, query tuning, code splitting and optimistic updates.

› Moved deployments to **Kubernetes** with automated rollouts, cutting deploy downtime from **5 minutes to under 2 seconds**.

› Hired and led a **5–7 engineer** team; introduced testing practices and code reviews that cut bugs per sprint from 20+ to 5; wrote the TDDs/RFCs.

Full Stack Developer · ICLEAR CCTV & Security System, Delhi (Remote) · May 2021 – Jan 2022 · Co-built the React frontend of a CCTV livestreaming app; repeat customers rose 30%.

PROJECTS & OPEN SOURCE

› FastBowlingLab · Founder, Aug 2026 – Present — AI fast-bowling coach built solo on web, iOS and Android: **MediaPipe** pose estimation on phone video (33 body landmarks, 19 measurements), LLM coaching, annotated video and adaptive training plans. **1,200+ real users**, 300+ monthly actives, ~20 paying customers, 2,500+ deliveries analysed. **Python (FastAPI)**, **Next.js**, **React Native (Expo)**, NestJS + tRPC, PostgreSQL, Inngest, Gemini, AWS ECS.

› livekit/components-js — Merged contribution to LiveKit's official React components, the library behind the WebRTC platform I built at Loktantra.

SKILLS

Languages:	**TypeScript**, JavaScript, Python, Go, SQL
AI / LLM:	**Generative AI**, **LLMs**, **AI agents**, **LangGraph**, LangChain, tool calling, structured outputs, **MCP servers**, **LLM evals**, guardrails, vector search, OpenAI, Anthropic Claude API, Gemini, computer vision (MediaPipe)
Frontend & Mobile:	**React**, **Next.js**, **React Native (Expo)**, Redux-Saga, WebRTC (LiveKit), Jest, Playwright
Backend & Data:	**Node.js**, **NestJS**, tRPC, **Python (FastAPI)**, GraphQL, WebSocket, **PostgreSQL**, pgvector, Elasticsearch, **Redis**, **Kafka**, Google Pub/Sub
Cloud & DevOps:	**AWS** (EKS, ECS, Lambda, RDS), **Kubernetes**, **Terraform**, Docker, GitHub Actions, OpenTelemetry
Security:	OAuth2, SSO, RBAC, SOC 2