Arjun Thakur

Principal / Lead AI Engineer | Architect & Fractional CTO | Founder | ex-Amazon, ex-Agoda

Available for: Lead / Principal AI Engineer | Architect | Fractional CTO | Full-time & freelance | Remote (global) | Based in Indore, India

+91-93401-58116 | thakurarjun247@gmail.com | arjunthakur.dev | linkedin.com/in/thakurarjun247 | github.com/thakurarjun247

SUMMARY

> AI engineer and principal-level architect who builds **production agentic systems end-to-end** -- LangGraph agent graphs, RAG pipelines, voice agents, and multi-tenant backends -- currently live with **5 paying CBSE schools** via **Yuvan**.

> **10+ years** shipping production systems at Amazon, Agoda, and high-growth startups across India, Southeast Asia, and the US; brings senior-engineering rigor to AI -- distributed systems, scale, security, cost engineering.

> Run **LangSmith evals, retrieval-precision metrics, and agent traces** before calling anything production-ready -- engineering discipline, not vibes-based prompting.

> Equally comfortable leading teams (15+ engineers, INR 30Cr P&L) and building solo as a senior freelancer / fractional CTO embedding agentic features inside someone else's stack.

CORE SKILLS

Agentic AI & LLMs: LangChain, LangGraph (stateful agent graphs), RAG, semantic FAQ caching, function/tool calling, prompt engineering, evals, agent observability (LangSmith), guardrails & verifier patterns, multi-agent orchestration

LLM Providers & Voice: OpenAI (GPT-4o / 4o-mini, Realtime API, Embeddings), Anthropic Claude, Gemini, OpenAI Whisper, WebRTC voice agents, streaming + interruption handling

Vector / Retrieval: pgvector, embeddings (text-embedding-3, Cohere, BGE), hybrid search, chunking strategies, cosine-similarity caching, retrieval evals

AI-Native Backend: Python 3.11+, FastAPI (async), Pydantic v2, asyncpg, pytest-asyncio, Java 21, Spring Boot, Spring Security, Hibernate JPA, Scala

Architecture: Microservices, distributed systems, event-driven (Kafka), HLD/LLD, design patterns, SOLID, clean code, API design (REST + OpenAPI), TDD

Data: PostgreSQL, pgvector, Cassandra, DynamoDB, Redis, MongoDB, SQL tuning, PgBouncer, Supabase (Auth + RLS)

Cloud & DevOps: AWS Certified Solutions Architect -- S3, ECS, Lambda, IAM, Serverless | Docker, Kubernetes, GitHub Actions, Jenkins, Railway, Vercel, Grafana, ELK

Frontend: Next.js 15 (App Router), TypeScript, JavaScript, HTML, CSS, React, Tailwind, shadcn/ui, KaTeX

Testing & QA: Test-Driven Development, JUnit, Mockito, pytest, pytest-asyncio, httpx

AI-Assisted Development: Claude Code, Cursor, GitHub Copilot, v0.dev -- shipping with AI as a daily collaborator

Leadership: Strategic planning, hiring & team growth, mentoring, curriculum design, stakeholder management, agile/scrum, cross-functional delivery, founder-mode product + GTM

Other: Data Structures & Algorithms, REST, system design

EXPERIENCE

Founder & AI Engineer | Yuvan (Voice-first AI Math Tutor)

Apr 2025 -- Present | Remote

Agentic AI | LangGraph | RAG | OpenAI Realtime | FastAPI | Next.js | Supabase | pgvector

> Built and shipped **Yuvan**, a voice-first AI tutor for CBSE Class 10 students -- onboarded **5 paying schools** as design partners on a B2B2C model (school recommends, parent pays); product currently live in MVP with paid pilots.

> Designed the full agentic architecture: a **LangGraph state machine** per doubt -- embed -> semantic FAQ-cache lookup -> RAG retrieval over NCERT chunks -> GPT-4o-mini generation -> math verifier -> topic auto-tagger -> persistence -- with an explicit re-explain loop on detected student confusion.

> Engineered a **WebRTC voice pipeline** on OpenAI Realtime (STT + TTS + VAD + barge-in interruption) with backend-minted ephemeral tokens; achieved sub-1.5s p95 first-audio latency in pilot sessions while keeping all reasoning, RAG, and verification server-side.

> Built a **RAG layer** on Postgres + pgvector (Supabase) -- ingestion, chunking, OpenAI embeddings, top-k retrieval, citation-grounded answers -- and a **global semantic FAQ cache** (cosine > = 0.92) targeting ~30% hit rate to push per-session OpenAI cost below the unit-economics ceiling.

> Implemented a **math-verifier safety net** that re-checks every numerical claim before the AI speaks it (<=1s overhead) -- the difference between a demo and a tutor that schools trust.

> Built the full multi-tenant product: 4 user roles (Student / Parent / School Admin / Super Admin), Supabase Auth + Postgres RLS for tenant isolation, parental-consent flow (DPDP / NCPCR-aligned), weekly parent reports, and a school engagement dashboard.

> Owned everything end-to-end: architecture, code, infra (Railway + Vercel + Supabase under $10/mo demo budget), LangSmith tracing, evals, school sales calls, parent-onboarding flows, pricing, and pilot success.

Director, Full-Stack Engineering | PhysicsWallah Nov 2024 -- Apr 2025 | Bengaluru / Indore

Java | Spring | Microservices | Python services | AWS | AI dev tooling

> Hired and led a team of **15 engineers** to build an in-house Learning Management System using Java, Spring, Hibernate, TypeScript, microservices and AWS.

> Stood up CI/CD on Jenkins + Docker, lifting deployment frequency by 15% and cutting release-night fire-drills.

> Introduced Python/FastAPI services for ML-adjacent workloads (analytics + content tagging) alongside the core Java stack, and rolled out AI-assisted development (Copilot, code review automation) across the team.

Independent AI Engineer & Builder | Agentic AI R&D Nov 2023 -- Nov 2024 | Remote

LangChain multi-agent patterns | voice pipeline prototyping | embeddings benchmarking

> Pivoted full-time into AI engineering and shipped an early AI math-teacher prototype to **900+ users** -- the user-validation work that proved demand and became Yuvan.

> Benchmarked **LangChain multi-agent patterns** (router, supervisor, plan-and-execute) on real student conversations to pick the architecture Yuvan now ships on.

> Prototyped four voice-agent stacks (OpenAI Realtime, Whisper + ElevenLabs, Whisper + Coqui, Gemini Live) and ran latency/cost shootouts -- the data behind picking Realtime + WebRTC.

> Cost-engineered embeddings choice across **OpenAI text-embedding-3, Cohere v3, BGE** for retrieval precision vs. INR-per-1M-tokens; documented trade-offs that drove the FAQ-cache design.

> Side build: production-grade microservices **flash-sale system** (Java 20, React, TypeScript) to keep distributed-systems chops sharp under high-concurrency load.

Senior Engineering Manager (Contract) | ForaySoft (client: Wayfair) Mar 2023 -- Oct 2023 | Remote

Java | Spring Security | Hibernate | TypeScript | PostgreSQL | Kubernetes | AWS

> Led 10 engineers on a fixed-term migration of Wayfair's seller-promotions microservice -- Java, Spring Security, Hibernate, PostgreSQL, Docker, AWS -- cutting system failures by 28%.

> Re-architected for scale on Kubernetes, supporting a 25% increase in transaction volume during peak e-commerce promotions.

Vice President, Engineering (Hands-on) | Nolan EduTech Oct 2021 -- Nov 2022 | Remote

Java | Spring | JPA | Microservices | AWS | TypeScript | Curriculum + GTM

> Launched a coding-bootcamp vertical from zero to **INR 30 crore (~$3.6M)** in revenue inside one year -- owned the curriculum, the platform, and the live teaching.

> Designed end-to-end backend + full-stack programs aligned to industry hiring; hit a **~95% placement rate** across cohorts in the thousands.

> Re-architected an internal HackerRank-style assessment platform on Java + Spring Boot + AWS for scale and reliability under load.

> Taught live online batches covering Java, backend systems, system design, and interview prep to thousands of learners.

Lead Developer | CoinSwitch Kuber Dec 2020 -- Oct 2021 | Remote

Python | FastAPI | Django | PostgreSQL | Redis | AWS | Microservices

> Migrated a monolithic crypto-trading app to microservices on **Python, FastAPI, and Django** on AWS -- the same async Python stack I now use for AI services.

> Rolled out Redis query caching to drop query latency by 30%, and tuned PostgreSQL with PgBouncer connection pooling, cutting query latency a further 40% and unblocking connection-saturation incidents.

> Held the system stable through a user surge from **600K to 10M** during a crypto bull run.

Software Development Engineer II | Amazon Jul 2019 -- Oct 2020 | Chennai

Java | DynamoDB | AWS | Payments / Digital Goods

> Hardened payment security against CSRF attacks, reducing impact by 19.7%.

> Architected a throttling system on DynamoDB + Java that cut coupon-code misuse by 15.3% in the first month post-launch.

> Integrated the Amazon Games payment flow into the Digital Payments Platform; improved payment-page accessibility for differently-abled users and search engines across geographies and devices.

Contract SDE | Porch.com, Domino, others Nov 2017 -- Mar 2019 | Seattle / Bengaluru

> Built a service from scratch integrating Porch.com's backend with Facebook Marketplace -- doubled service-quote requests and lifted revenue 13% by tapping ~200M FB users.

> Optimized MongoDB analytics sorting via Python tooling for Domino, improving response time 30%.

Senior Software Developer | Agoda.com Aug 2016 -- Aug 2017 | Bangkok

Scala | Cassandra | Kafka | Distributed systems

> Engineered an in-house A/B testing pipeline processing millions of messages on Scala + Cassandra + Kafka -- cut time-to-insight for product managers by 22%.

> Wrote shared Scala libraries that grew tracked-user coverage by 10%.

Software Engineer | Hadoolytics Technologies Apr 2015 -- Aug 2016 | Gurugram

> Architected a real-estate platform from scratch on Java, Spring, AWS, and Cassandra.

> Integrated Swagger for auto-generated API docs, saving ~20% of development time per cycle.

EDUCATION & CREDENTIALS

M.Tech, Computer Science & Engineering | NIT Jalandhar -- CGPA 7.5/10. Taught and mentored thousands of undergraduate students as a teaching assistant / instructor -- the first chapter of a teaching thread that later became Nolan EduTech and PhysicsWallah.

B.E., Computer Science & Engineering

AWS Certified Solutions Architect

SELECTED PROJECT HIGHLIGHTS

Yuvan -- voice-first AI tutor; LangGraph agent | RAG on pgvector | OpenAI Realtime WebRTC | math verifier | semantic FAQ cache | multi-tenant Supabase RLS | 5 paying schools. *Architecture deep-dive and code samples available on request.*

Flash-sale system -- production-grade microservices simulating real concurrency, queueing, and back-pressure on Java 20 + React + TypeScript.

Online assessment platform -- HackerRank-style code-execution and grading on Java + Spring Boot + AWS, scaled for thousands of concurrent test-takers.