Arjun Thakur

Lead AI Engineer | Agentic Systems | RAG | LLM | Founder | ex-Amazon, ex-Agoda

Available for: AI engineering contracts & freelance | Remote (global) | Indore, India

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

SUMMARY

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

> **10+ years** at Amazon, Agoda, and high-growth startups; 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.

CORE AI SKILLS

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

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

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

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

Data & Infra: PostgreSQL, pgvector, Cassandra, DynamoDB, Redis, Supabase (Auth + RLS), AWS (S3, ECS, Lambda), Docker, Kubernetes, Railway, Vercel

AI-Assisted Dev: Claude Code, Cursor, GitHub Copilot, v0.dev

SELECTED AI WORK

Founder & AI Engineer | Yuvan (Voice-first AI Math Tutor) Apr 2025 -- Present | Remote

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

> Shipped **Yuvan**, a voice-first AI tutor for CBSE Class 10 -- onboarded **5 paying schools** as design partners on a B2B2C model; product live in MVP with paid pilots.

> Designed the agentic core: 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 a re-explain loop on detected confusion.

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

> Built a **RAG layer** on Postgres + pgvector and a **global semantic FAQ cache** (cosine > = 0.92) targeting ~30% hit rate to keep per-session OpenAI cost inside unit economics.

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

> Multi-tenant from day one: 4 user roles, Supabase Auth + Postgres RLS, parental-consent flow (DPDP / NCPCR), weekly parent reports, school engagement dashboard.

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

> Shipped an early AI math-teacher prototype to **900+ users** -- validation that became Yuvan.

> Benchmarked LangChain multi-agent patterns (router / supervisor / plan-and-execute), four voice-agent stacks, and three embeddings models -- the data behind Yuvan's stack picks.

SENIOR BACKEND CREDIBILITY (10+ YRS)

Director, Full-Stack | PhysicsWallah (Nov 2024 -- Apr 2025) -- led 15-engineer LMS rebuild; introduced Python/FastAPI services + AI dev tooling across the team.

Sr. Eng. Manager (Contract) | Wayfair via ForaySoft (Mar -- Oct 2023) -- led 10 engineers; cut system failures 28%, +25% transaction volume on Kubernetes.

VP Engineering | Nolan EduTech (Oct 2021 -- Nov 2022) -- launched coding-bootcamp vertical to INR 30Cr (~$3.6M) revenue in one year; 95% placement rate.

Lead Developer | CoinSwitch Kuber (Dec 2020 -- Oct 2021) -- Python / FastAPI / Django microservices; held system stable through 600K -> 10M user surge.

SDE-II | Amazon (Jul 2019 -- Oct 2020) -- payments / digital goods; throttling on DynamoDB cut coupon misuse 15.3%; CSRF hardening cut impact 19.7%.

Sr. Software Developer | Agoda.com, Bangkok (Aug 2016 -- Aug 2017) -- Scala + Cassandra + Kafka A/B testing pipeline at scale.

Earlier: Porch.com, Domino, Hadoolytics -- full timeline on long-form resume / LinkedIn.

EDUCATION & CREDENTIALS

M.Tech, CSE | NIT Jalandhar (CGPA 7.5) -- taught thousands of undergrads as TA | **B.E., CSE** | **AWS Certified Solutions Architect**