

KUNAL ROY CHOUDHURY

[Email](#) | [LinkedIn](#) | [GitHub](#) | [Portfolio](#) | +91-9954636583

EXPERIENCE

Lead Software Engineer

March 2026 – Present

FinStocks AI

Bengaluru, Karnataka

- Led a 3-engineer team building a **LangGraph** agent for live Indian-market chat, charting, research and orders; collapsed a 3-stage router into a single **tool-call node** that also emits the typed research plan, removing one serial model call per turn.
- Cut chart latency with **speculative lane racing**: cheap chart lane runs in parallel with routing, wins only on a valid op batch and cancels the loser; money-moving actions stay sequential and are asserted at import.
- Built a **grounded answer pipeline** with hard timeouts and a validator that checks every numeral/citation against live evidence; models stay off the order path (**place_order** takes no args and rejects mismatched confirmations).
- Shipped an **LLM eval & observability platform**: 73 versioned prompts, 18 model configs, per-turn lineage (tokens, cost, TTFT), A/B seams, and a 3-state CI gate (pass / regression / inconclusive) fed by production failures.
- Hardened the **GCP Cloud Run** backend: Postgres job queue (**SKIP LOCKED**, visibility leases), connection-budgeted admission control, and a single tick stream fanned to all WebSockets; cut market-data queries from **13.6 s to 48–135 ms**.

Software Developer

April 2025 – February 2026

Indian Kanoon

Bengaluru, India

- Sole engineer for [Prism](#), Indian Kanoon's AI legal-research suite; designed, built, and launched the full product (Django + PostgreSQL + Redis) into a live platform serving **1.5M+** users with 10+ LLM-powered tools.
- Engineered a production RAG pipeline (PostgreSQL full-text search, query expansion, multi-level caching, index tuning) that cut median latency by **40%** and raised grounded query accuracy to **92%** across millions of judgments.
- Architected multi-tool LLM workflows (retrieval → reasoning → generation) powering specialized agents for legal Q&A, document drafting, case prediction, and counter-argument generation, with citation grounding and evaluation loops for production reliability.

Software Engineer Intern

September 2024 – April 2025

Chargebee

Chennai, India

- Migrated **500K+** **subscriptions** with zero customer-facing downtime using a migration framework built with **JOJO** schedulers and **Vue.js** validation tooling.

Software Engineer Intern

January 2024 – September 2024

AiDash

Bengaluru, India

- Contributed to a **60% faster** release cycle by extracting services from a **2M+ LOC monolith** using **Spring Boot** and hexagonal architecture; optimized SQL pagination over 100K+ record tables.

PROJECTS

Codebase Indexer | *Rust, Tree-sitter, Tantivy, ONNX, HNSW, RRF Fusion*

[code](#) | [article](#)

- Indexed **50K files in 21 seconds** with a local **Rust** hybrid search engine combining Tree-sitter chunking, Tantivy BM25, and HNSW retrieval via RRF; used **INT8 ONNX** models for local inference.

Gostman | *Go, React, Wails, REST, GraphQL, WebSockets*

[live](#) | [code](#)

- Achieved **<1s cold launch** and **90% lower memory** than Electron equivalents with a privacy-first desktop API client featuring local-only storage and request chaining across REST, GraphQL, and WebSocket protocols.

TECHNICAL SKILLS

Languages: Go, Rust, Python, Java, TypeScript, SQL

Systems & Performance: concurrency, cgo/FFI, profiling, caching, batching, backpressure, circuit breakers

AI / LLM: LangGraph, RAG, tool calling, vector search (HNSW), ONNX, Anthropic & OpenAI APIs

Backend: Django, FastAPI, Spring Boot, gRPC, REST, GraphQL, WebSockets, WebRTC, Celery

Data & Infra: PostgreSQL, Redis, Elasticsearch, AWS, GCP, Docker, Kubernetes, CI/CD

Frontend: React, React Native (Expo), Vue.js

EDUCATION

SRM Institute of Science and Technology

B.Tech, Computer Science and Engineering — CGPA: 9.24 / 10

Kattankulathur, India

August 2020 – September 2024