

Kartik Maurya

Mumbai, India | [8879697133](https://github.com/kartikmaurya18) | kartikmaurya18@gmail.com | [GitHub](#) | [LinkedIn](#)

SUMMARY

AI/LLM Engineer with production backend experience across banking, fintech, and healthtech. Architected a RAG platform with LangGraph multi-agent orchestration and multi-LLM routing that cut inference latency 62% and API costs 35%, with guardrails blocking 95%+ of unsafe requests. Backed by production depth: 100+ REST APIs in Java/Spring Boot and a monolith-to-microservices migration for a national banking platform.

EXPERIENCE

In-Solutions Global | *Software Development Engineer*

Nov 2025 – Present

- Led the migration of PNB Genie from a monolithic Apache Struts architecture to Spring Boot microservices across App, User, SI, and Admin services, improving deployment velocity and long-term maintainability.
- Designed and shipped 100+ RESTful APIs in Java, Spring Boot, and Oracle SQL (JPA/Hibernate) for Punjab National Bank and Saraswat Bank, powering core banking transaction workflows under production reliability and security constraints.
- Resolved SIT and UAT defects across core banking modules and executed production releases via WinSCP, validating builds with Maven and Postman regression suites.

Yumedics Labs | *AI Engineer intern*

Jun 2025 – Oct 2025

- Built an AI-powered operations dashboard for Yumedics and Yuderma, automating Gmail and WhatsApp invoice ingestion and processing for a 15-member field team, eliminating manual data entry from the invoicing workflow.
- Implemented Role-Based Access Control (RBAC) and real-time activity tracking; partnered directly with founders to convert business requirements into production-ready automation workflows for the field team.

PROJECTS

Enterprise AI Copilot Platform | *Python, LangGraph, Qdrant, Redis, Docker*

[GitHub](#)

- Architected a production-grade AI Copilot with RAG, LangGraph multi-agent orchestration, and multi-LLM routing via an AI Gateway with semantic caching, cutting inference latency 62%, API costs 35%, and boosting throughput 2.5x.
- Implemented AI guardrails with LLMops monitoring, PII detection, and prompt-injection defense, blocking 95%+ of unsafe/malicious requests before reaching the model layer.

OptionOS | *FastAPI, React.js, WebSockets, IndexedDB*

[GitHub](#)

- Built a real-time BTC options analytics platform with a dual-WebSocket architecture and strike-indexed in-memory caching, streaming live options-chain and straddle data with zero data gaps.
- Built an OHLC data pipeline with IndexedDB persistence, enabling low-latency chart rendering directly in the browser.

Full Stack AI Chatbot | *FastAPI, React.js, Redis Streams, GPT-J*

[GitHub](#)

- Built a real-time AI chatbot using FastAPI, React, WebSockets, Redis Streams, and GPT-J-6B; implemented scalable client-server architecture with asynchronous message processing.
- Designed a Redis Streams producer-consumer pipeline and integrated the Hugging Face Inference API for LLM inference and persistent chat history, optimizing response latency.

EDUCATION

Bachelor of Engineering (B.E.) – Information Technology: University of Mumbai - TCET

CGPA: 8.37 | 2026

SKILLS

AI & LLM Engineering: RAG, LLMops, Multi-Agent Orchestration (LangGraph), Multi-LLM Routing, Semantic Caching, AI Guardrails, PII Detection, Prompt-Injection Defense, Hugging Face

Data, Vector Stores & Messaging: Qdrant (Vector DB), Redis, Redis Streams, Apache Kafka, Oracle SQL, PostgreSQL, IndexedDB

Languages: Python, Java, JavaScript, SQL (Oracle)

Backend & Architecture: FastAPI, Spring Boot, JPA/Hibernate, Django, Node.js, Apache Struts, Microservices Architecture, REST API Design, WebSockets, RBAC

DevOps & Cloud: Docker, Jenkins, CI/CD Pipelines, AWS, Git, Maven, Postman