

Yukti

Software Engineer | Backend (Java · Spring Boot · Python) | Distributed Systems | AI & LLM Integration (RAG)
Bengaluru, India • Open to full-time & contract roles • yukwork480@gmail.com
linkedin.com/in/yukti-6b4512254 • github.com/Yukti007

Professional Summary

Software Engineer with 1+ year at Visa building backend services for its risk and settlement platform in Java and Spring Boot, and integrating LLMs into internal tools used by business teams. Shipped an event-driven alerting service (~4 minutes faster), a RAG pipeline that cut LLM context by ~96%, and a 10,000+ record production data migration. Also builds AI systems from scratch and contributes C++/MQTT firmware to an open-source IoT device.

Technical Skills

Languages: Java, Python, SQL, TypeScript, JavaScript, C++

Backend & Distributed Systems: Spring Boot, FastAPI, REST APIs, Microservices, Event-Driven Architecture, Asynchronous Messaging (IBM MQ, MQTT), Object-Oriented Design, Design Patterns

Applied AI: LLM Integration, Retrieval-Augmented Generation (RAG), Semantic Search, Tool/Function Calling, Structured Outputs, Prompt Engineering

Data: MySQL, SQL Server, SQLite, SQLAlchemy, Microsoft Dataverse, Relational Data Modeling, Data Migration

Cloud & DevOps: AWS (EC2, Lambda, S3, IAM), Docker, GitHub Actions (CI/CD), Linux, Git, Maven

Also: Unit Testing, Agile/Scrum, Technical Documentation, Power Automate, Power Apps

Professional Experience

Software Engineer

Visa

Bengaluru, India

Jun 2025 – Aug 2026

- Built and shipped an event-driven alert processing service (Java, Spring Boot, IBM MQ) that decoupled alert generation from the core settlement platform, cutting alert processing latency by ~4 minutes and improving scalability and reliability.
- Designed backend services and secure REST API integrations for an AI-assisted conversational feature in REACT, Visa's risk-exposure tool, so CSR and business teams could retrieve settlement exposure, client limits and operational insights in natural language.
- Engineered a Retrieval-Augmented Generation (RAG) pipeline for duplicate-idea detection using Dataverse AI Native Search and Power Automate, reducing LLM context by ~96% (120K to 3–5K tokens) and lowering inference cost.
- Led the migration of a legacy SharePoint ideation platform to Microsoft Dataverse: redesigned the relational data model, migrated 10,000+ production records, and wrote documentation for deployment and maintenance.
- Wrote unit tests for backend services and worked in Agile sprints with product managers, taking features from requirements through design, testing, deployment and support.

Open Source Contributor – Embedded Firmware

Public Invention (Krake)

Remote

Nov 2025 – Present

- Merged 20+ pull requests into open-source C++ firmware for Krake, a wireless ESP32 alarm device, working on MQTT messaging, WiFi/TCP-IP connection handling and alarm features.
- Debugged and tested changes on real hardware, and worked through design trade-offs with maintainers in code review.

Projects

VoiceAgent – Multilingual Realtime AI Voice Agent | *Python, FastAPI, LiveKit, SQLAlchemy, OpenAI*

- Built a realtime voice agent that converses in English, Hindi and Hinglish and books appointments through LLM tool calls backed by a relational database, with support for mid-sentence interruptions.
- Split the system into independent API and agent-worker processes, with a provider-agnostic LLM layer, schema-validated post-call summaries and per-stage latency monitoring.

Bouncer – Guardrail Middleware for LLM APIs | *Python, FastAPI, Pydantic*

- Built async middleware that redacts PII and blocks prompt-injection patterns on inbound requests, and enforces strict schemas on LLM responses, returning structured errors instead of malformed output.

Education

RNS Institute of Technology

Bachelor of Engineering, Computer Science & Engineering (AI & ML)

Bengaluru, India

2021 – 2025 | CGPA: 8.9/10