

Kalyan Mudumby

Boston, MA | 857-295-3444 | mudumby.k@northeastern.edu
linkedin.com/in/kalyan-mudumby | github.com/theinhumaneme | kalyanmudumby.com

Backend Engineering | Infrastructure & Developer Tools | Applied AI

Education

Northeastern University, Boston, MA

September 2025 - May 2027

Master of Science in Computer Science

- Coursework (completed or in progress): CS5800 Algorithms, CS5610 Web Development, CS5010 Programming Design Paradigm, CS5200 Database Management Systems, CS6140 Machine Learning, CS5330 Pattern Recognition and Computer Vision (in progress).

Geethanjali College of Engineering and Technology, Hyderabad, India

August 2019 - May 2023

Bachelor's Degree in Electronics and Communication Engineering

Technical Skills

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

Backend: Spring Boot, FastAPI, Flask, Axum, Tokio, REST APIs, WebSockets

Frontend & Visualization: React, Angular, Next.js, D3.js, MapLibre GL

Databases & Search: PostgreSQL, Redis, MongoDB, Meilisearch

Cloud & DevOps: Azure, Cloudflare, Docker, Docker Compose, Linux, Nginx, GitLab CI, GitHub Actions, Git

Testing: JUnit 5, Testcontainers, Mockito, pytest, Vitest, Playwright

Observability: OpenTelemetry, Grafana, Tempo, Loki, Mimir, Prometheus

AI / LLM: LangChain, LangGraph, vLLM, LM Studio, OpenAI & Gemini APIs, MCP, RAG

Experience

Northeastern University, Boston, MA

September 2026 - Present

Graduate Teaching Assistant, CS4530: Fundamentals of Software Engineering

- Support **250+ students weekly** in labs and office hours on React, React Hooks, and component-based architecture.
- Mentor 60+ students on MVC architecture, Vitest unit tests, Playwright usability tests, and mocking strategies.

Nebulogic Technologies, Hyderabad, India

June 2023 - July 2025

Software Engineer

- Cut core CRM cold-start time **100x** (20s to 200ms) by batching cache-warmup writes with Redis pipelining. Redis Pub/Sub propagated configuration changes across running instances without restarts.
- Reduced page load from **30s to 7s** with Brotli/Gzip compression and a 60% smaller JavaScript bundle. Cloudflare CDN and Nginx caching cut origin load by up to 70%, with an 80% static-asset cache-hit ratio.
- Built SMART AI Chat, a RAG chatbot answering Fort Wayne residents' questions about city services using OpenAI and LangChain. Cached repeated questions and expanded its database around the information residents requested most.
- Made slow requests traceable from dashboards to logs with OpenTelemetry, Grafana, Tempo, Loki, and Mimir. Request-rate, error-rate, and latency dashboards identified slow services in a request chain.
- Fixed a Pydantic validation error in LangChain's GPTRCache integration for an unhandled ChatGeneration type; PR #19427 merged upstream.

Nebulogic Technologies, Hyderabad, India

June 2022 - June 2023

Software Engineer Intern

- Standardized builds across **40+ services** with Docker multi-stage builds on GitLab CI for reproducible deployments.
- Upgraded GitLab from v14 to v16 while preserving existing repositories and build pipelines. Tuned the instance for performance and reliability. Added disaster recovery so the team could restore the GitLab server from backups.

Additional Nebulogic work: Integrated Esri ArcGIS layers for data collection in Fort Wayne's 311 product. Combined the data with other city providers to give citizens precise trash-collection locations.

Awards & Certifications

Best Podman Container Implementation & People's Choice | Red Hat/IBM, OAAM

March 2026

Best Employee of the Month | Nebulogic

May 2024; February, September, October 2023

AWS Certified AI Practitioner (August 2026) | Microsoft Certified: Azure Fundamentals (AZ-900) (May 2023)

Selected Projects

Documan

June 2020 - August 2026

Python, Flask (historical); Java 25, Spring Boot, Angular 21, PostgreSQL (rewrite)

github.com/theinhumaneme/documan-public-release | Live site

- Served **4,000+ students** through a college digital library on Python and Flask. The historical service, ending May 2025, delivered 200,000+ files to 1,200+ active users through a web app and Telegram bot at 99%+ uptime.
- The rewrite enforces role-based access control in four tiers, restricting maintainers to their assigned department, year, and semester. Authorization tests check every data-writing endpoint for a role rule.
- Document uploads in the rewrite remain available during search outages. Committed library changes stay queued for indexing; rolled-back changes are never indexed.

Hikari

September 2024 - April 2025

Rust, Tokio, Axum, PostgreSQL, Docker Compose, WebSockets

github.com/theinhumaneme/hikari

- Hikari replaces separate container configurations on each machine with one JSON file describing what runs where. It generates container stacks, and each machine brings itself in line with its assigned configuration.
- Configuration changes redeploy only affected stacks, keeping unrelated services running. Agents reconnect automatically after disconnection. Configuration payloads stay encrypted.

Patternsmith

June 2026

Python, LangGraph, LangChain, Hugo, vLLM, LM Studio

github.com/theinhumaneme/patternsmith | Live site

- Patternsmith's AI agents turn a software design pattern or code smell into a tutorial with prose, code, and a diagram. Validation retries malformed responses and diagrams. A failed topic does not stop the batch.
- The same pipeline uses hosted APIs or local models served through vLLM and LM Studio. LM Studio runs use quantized models and multi-token prediction. Offline mock-model tests exercise the pipeline without API calls.

Activity Visualizer

TypeScript, React, MapLibre GL, Web Workers, WebCodecs, Cloudflare Workers

activity-visualizer.distro.dev

- Activity Visualizer turns GPX and TCX fitness exports into performance analysis, 3D route flyovers, and shareable videos. Files are analyzed in the browser. Privacy zones hide route locations near the start and finish on maps and exports.
- Background workers parse bulk imports and recover from crashes. A map-tile proxy keeps provider keys server-side. Video exports retry unavailable tiles and report incomplete frames.

Training Story (Hevy Visualizer)

JavaScript, React, D3.js, GSAP, Chart.js, Playwright, Cloudflare Pages

github.com/theinhumaneme/hevy-visualizer | Live site

- Training Story turns workout CSV exports into interactive dashboards and an animated training story. A shared analytics engine cleans and deduplicates records without accounts or workout-data uploads.
- Calculation tests compare training metrics with hand-computed results. Playwright tests exercise the production build across app views, mobile layouts, and reduced-motion settings.

OAAM (Open Agent Applet Marketplace)

March 2026

React, TypeScript, Node.js, Express, Podman AI Lab, MCP

- OAAM keeps document processing on the user's machine: a local browser app extracts structured data without a cloud inference service. A vision model reads an invoice and returns a checked record.
- A tool registry exposes AI tools through one interface. An optional MCP extension sends extracted results to Google Sheets.