

Laxman K R

Chennai, India | +91 7448657412 | krlaxman03@gmail.com

linkedin.com/in/laxmanramesh | github.com/thaletto | laxmankr.vercel.app

SUMMARY

Generative AI Engineer with production experience building Python and FastAPI systems for agentic incident automation, RAG, and intelligent workflows. Experienced in designing end-to-end retrieval pipelines spanning document ingestion, embeddings, hybrid and vector search, MCP-enabled tool calling, and REST integrations with Qdrant and cloud-native services. Builds full-stack AI products and open-source developer tooling using TypeScript and Python, while leveraging Claude Code, OpenAI Codex, and GitHub Copilot for AI-assisted implementation, debugging, refactoring, and development workflows.

TECHNICAL SKILLS

Languages: Python, TypeScript, SQL, Go, JavaScript

Generative AI: LLMs, Prompt Engineering, RAG, Text-to-SQL, Embeddings, Vector Search, Agent Orchestration, Tool Calling, MCP

AI Frameworks & Tools: Claude Code, OpenAI Codex, Gemini, AutoGen, LangGraph, LangChain, Google ADK, AI SDK, Qdrant, Pinecone

Backend & Data: FastAPI, Flask, REST APIs, PostgreSQL, MSSQL, Redis, MongoDB, SQLite, Node.js, Express

Frontend & Platform: Next.js, React, TanStack Start, Effect TS, AWS Lambda, Google Cloud, Docker, Kubernetes, Cloudflare, Grafana Loki, GitLab CI/CD, CI/CD

WORK EXPERIENCE

Tata Consultancy Services

Nov 2024 – Present

AI Engineer

Chennai, TN

- Built Python-based production incident automation that integrates Grafana Loki anomaly signals with ServiceNow REST APIs, cutting operational **MTTR by 63%**.
- Reduced average incident resolution time from 25-30 minutes to 6-8 minutes by integrating a Qdrant-backed RAG context-engineering process into a multi-agent incident-resolution workflow.
- Reduced manually resolved incidents from 20 to 3-4 per month by shipping a Next.js dashboard with human-resolved, approval-gated, and automated AI-agent paths.
- Led a migration of **50+ modules** from a monolithic Next.js app, halving build time and raising deployment throughput 2.5x.
- Partnered with the Google Cloud team to ship a platform representing TCS at **Google Cloud Next 2026**, showcasing **120+ AI offerings** to Google and TCS sales teams.
- Owned CI/CD across infrastructure, sales, and business-analysis teams, supporting **25+ production deploys/month** with zero critical rollbacks.

PROJECTS

PROFESSIONAL

Agentic AMS Platform

Feb 2025 – Jan 2026

AutoGen, FastAPI, MCP, Qdrant, Grafana Loki, ServiceNow

Case Study

- Designed a FastAPI and Microsoft AutoGen multi-agent system that detects, diagnoses, and remediates production failures through MCP-enabled tool calling.
- Built an observability agent that ingests Grafana Loki alerts over WebSockets, deduplicates ServiceNow tickets, and runs document ingestion, chunking, embeddings, and semantic retrieval through a Qdrant RAG index.
- Implemented a Planning–Executor–Validator agent workflow with **human-in-the-loop** approval for infrastructure rollbacks, structured validation, and automated pod health checks.

Google GTM Portal

Jan 2026 – Apr 2026

Next.js, FastAPI, GCS, Google Cloud

Case Study

- Architected a typed FastAPI and Google Cloud Storage catalog backend with a Next.js frontend, delivering sub-second loads and a 96% Lighthouse score.
- Implemented hybrid keyword and vector retrieval to improve discovery across a catalog of 120+ AI offerings.

OPEN SOURCE

Cortex

Apr 2026 – May 2026

Effect TS, TypeScript, Pinecone, Qdrant

Docs

- Authored an Effect-native vector database ORM with structured concurrency, typed dependency injection, and compositional error handling.
- Shipped provider adapters (Zvec, PageIndex) so multi-tenant RAG pipelines stay portable across vector backends.
- Raised batch insert throughput to **3.2x baseline** using non-blocking batch stream pipelines on Pinecone and Qdrant.

Ascendant

Sep 2025 – Apr 2026

Python, Effect TS, PyPI, npm, AI Skills, AI Tools

Docs

- Built Ascendant, an astrology product delivered as dependency-free core libraries: a Python package on PyPI and an Effect TS-native npm package, providing high-precision ephemeris and planetary-positioning calculations.
- Published typed SDK APIs, AI agent skills, and a tools plugin (**npx skills**) for 16 divisional coordinate charts and planetary period timelines.

PERSONAL

Kite

Mar 2026 – Apr 2026

AG-UI, AI SDK, TanStack Start, React 19

Demo

- Built an AI assistant that streams UI over the AG-UI protocol, orchestrated by an AI SDK agent in TanStack Start, rendering interactive components alongside text responses.
- Defined server-driven UI schemas with graceful fallbacks, delivering flicker-free dynamic widgets on interactive analytics views.

Retail Platform

Jul 2024 – Jun 2026

Go, Next.js, AWS Lambda, Cloudflare D1, Workers

GitHub

- Designed 7 event-driven Go microservices on AWS Lambda (search, recommendations, notifications, orders) behind a Next.js admin/CMS for 200+ products.
- Used Cloudflare Edge caching, D1 databases, and Workers to absorb **10k req/min spikes**, cutting **p99 API latency from 450ms to 85ms**.

ACHIEVEMENTS

- Vercel OSS Contributor (contributed to ai-elements, a repo of UI components for building AI web apps).
- Open Source Impact: Cortex, an Effect TS vector ORM, reached **480 npm downloads**; astro-ascendant ephemeris software is published on PyPI.
- TCS Special Recognition Award (2025) for standout performance in the company-wide AI Ignite Hackathon.

EDUCATION

Amrita Vishwa Vidyapeetham

Sep 2020 – May 2024

B.Tech, Computer Science and Engineering

Coimbatore, TN

- Research: complex/irregular lung nodule detection from CT scans (medical image processing).