

Sreedev TS

+91 89437 50702 | xreedev@gmail.com | linkedin.com/in/sreedevts | github.com/xreedev

SUMMARY

AI/ML backend engineer with 2+ years shipping production GenAI systems for European hospitals under **GDPR**. Deployed an MCP server across multiple hospitals exposing 15+ AI tools, and cut LLM token cost 45% via RAG-based tool selection, model routing, and prompt caching.

EDUCATION

Marian Engineering College, APJKTU
B.Tech in Computer Science Engineering – CGPA: 7.71

Dec. 2020 – June 2024
Thiruvananthapuram

EXPERIENCE

Software Engineer (AI/ML Backend)

Jul 2024 – Present

Simplogics

Technopark, Thiruvananthapuram

MOM AI Secure Data Analysis Tool

- Built and deployed a production **Model Context Protocol (MCP)** server across multiple hospitals, exposing existing backend APIs as **15+ AI tools** via **FastAPI**, orchestrating multi-step tool execution with **LangGraph**.
- Reduced LLM token cost **45%** while improving response latency, using **RAG**-based tool selection over **ChromaDB**, intelligent model routing, and system prompt caching.
- Shipped an LLM analytics platform for data retrieval, visualisations, and automated Excel reporting, with a Deep Research mode selecting SQL templates dynamically; **15,000+ queries** in production.

MOM AI Secure NLP Chatbot

- Engineered a **self-learning Rasa** NLP platform that builds its own training data from live user queries and learns spelling variants via fuzzy matching, with automated retraining and **30+ custom Python components**; served **125,000+ SQL queries** in production.
- Built **NER** with custom **spaCy** models and an annotation pipeline generating training data directly from Excel.

Platform & Infrastructure

- Built an **evaluation harness** over a labelled set of production queries, scoring entity-level **precision/recall** for NER and **tool-selection accuracy** for MCP calls; run on each release to catch regressions.
- Built a pseudonymisation module for **PHI** over French hospital activity and billing data (**GHM/GHS, CCAM**, emergency-care pathways), preserving analytic context across LLM data paths under **GDPR**.
- Designed an **event-driven** architecture on **Redis**, moving Excel generation, **ETL**, and AI inference to async consumers off the request path; built the **Python/Node.js** microservices over **MSSQL, Jenkins CI/CD** with **Nexus**, and **Grafana/Prometheus/Loki** observability.

IAIFI (NSF AI Institute) Virtual Summer School

- Studied **symbolic regression** and **PySR** under the creator of PySR, and ran an equation-discovery task on real datasets with allocated **HPC** compute, recovering interpretable closed-form models.

PROJECTS

NyayBandh | *Sarvam Epoch Buildathon (GrowthX × Sarvam AI)*

Jul 2026

- Selected as **1 of 200 builders from 3,000+ applicants** for the invite-only buildathon.
- Built a **live negotiation mediator** on the **Sarvam AI stack**, orchestrated with **LangGraph**, detecting misunderstandings between parties in real time before they escalate into disputes; shipped end-to-end within the 8-hour window.

Board Flex | *Winner, Internal Hackathon — Simplogics*

- 1st of 15 teams** in a **16-hour** hackathon: event scheduler and live scoring platform in **Python/FastAPI/PostgreSQL, OpenAI API**, on **AWS EC2**.

TECHNICAL SKILLS

Languages: Python, SQL, JavaScript (Node.js), Java, C

AI & LLMs: RAG, MCP, LangGraph, Prompt Caching, LLM Evals (Ragas, Langfuse), NLP, NER, Embeddings, Vector Search, ChromaDB, OpenAI API, Rasa, spaCy, PySR

Backend & Data: FastAPI, Node.js, Event-Driven Architecture, Redis, Microservices, MSSQL, PostgreSQL, ETL, Pandas

DevOps & Cloud: AWS (EC2), Azure (Container Apps, AI Foundry), Jenkins, Nexus, Grafana, Prometheus, Loki, Git; Claude Code, Cursor