Sai Kiran Vepamani

Lead Gen AI Engineer

+91 9493922218 | vepamanisaikiran@gmail.com | vepamanisaikiran.github.io/portfolio | vepamanisaikiran | saikiranvepamani

Skills

Languages	Python · Java · C++/C · Rust · TypeScript · Node.js · SQL · Bash
AI/GenAI	LLM Fine-Tuning (QLoRA/LoRA/PEFT/RLHF) · RAG/PathRAG · Prompt Engineering · Agentic AI · LangChain · LangSmith · LangGraph · DSPy · Agno
Frameworks	FastAPI · Django · Spring Boot · Next.js · React · Angular · PyTorch · TensorFlow · Transformers · Scikit-learn
Cloud/Tools	AWS · Docker · Kubernetes · MCP · A2A · gRPC · Protobuf · Git · CI/CD · Playwright · Jenkins

About me

Lead Engineer - Gen AI at Bosch with 6+ years in Generative AI, LLM fine-tuning, FuSA AI, and cloud-native development. Led teams building RAG platforms, HARA AI for automotive safety, and multi-tenant SaaS. Patent co-author, hackathon winner (Bosch AWS 2025, AppsForBharat 2025), and Bharat Mobility Expo 2025 presenter.

Work Experience

Bosch Global Software Technologies
Bangalore, India

Lead Engineer - Gen AI
Mar. 2023 - Current

- Leading a cross-functional team of 8+ engineers across Gen AI, automotive middleware, and cloud platform initiatives.
- Built HARA AI for Hazard Analysis and Risk Assessment, automating ASIL classifications per ISO 26262.
- Architected BRICK, a document analysis platform with PathRAG for high-precision retrieval from unstructured PDFs.
- Won Bosch AWS Hackathon 2025 for ”Project IQ,” achieving 90% accuracy extracting competencies from SRS documents.
- Performed PEFT on Qwen 7B+ models using QLoRA/LoRA with 4-bit quantization on limited hardware.
- Architected Serverless Multi-Tenant SaaS platform using AWS CDK, reducing onboarding time by 90%.
- Developed CAN-to-VSS protocol mapper and optimized KUKSA.val databroker in C++/Rust for low-latency signal orchestration.

HashedIn by Deloitte
Bangalore, India

SDE - II
Aug. 2021 - Mar. 2023

- Built cloud-native content normalization platform for Thomson Reuters legal domain using AWS Lambda, API Gateway, DynamoDB, and CloudFormation.
- Developed annotation component in Angular that reduced document content processing time from 7 days to 4 hours.
- Collaborated with US legal domain stakeholders for requirement gathering and defining technical user stories.

Zapcom Solutions
Bangalore, India

Software Engineer
Feb. 2020 - July 2021

- Built REST APIs using Python/Django DRF with Celery async processing and NLP scoring using spaCy.
- Built Content Management service interface handling 1000+ client devices in WSO2.
- Wrote automation test scripts in Python using Selenium, reducing manual testing effort by 35%.

Zapcom Solutions
Bangalore, India

Full Stack Intern
Jan. 2019 - Jan. 2020

Achievements & Awards

2025	**Patent,** Co-Author -- "A control unit for detunneling of data from an ethernet frame" (No: 202541072960)	*India*
2025	**Winner,** Bosch AWS Hackathon 2025 -- "Project IQ," AI-powered training & knowledge management system	*Bosch*
2025	**Presenter,** Bharat Mobility Expo 2025 -- Showcased Vehicle Assistant AI that generates SDV apps deployable on HMI and HPCs	*Bosch*

Education

JNTUA(Jawaharlal Nehru Technological University Anantapuramu)
Ananthapur, India

B.Tech in Computer Science and Engineering with 7.74 GPA
Aug. 2015 - May. 2019