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 · MySQL · PostgreSQL · Bash
AI/GenAI	LLM Fine-Tuning (QLoRA/LoRA/PEFT/RLHF) · RAG/PathRAG · Prompt Engineering · Agentic AI · DSPy · Agno · LangChain · LangSmith · LangGraph · Hugging Face · Vector DBs (FAISS/ChromaDB)
ML/Data	PyTorch · Tensorflow · Transformers · Scikit-learn · OpenCV · Pandas · Numpy · Optuna · MLflow · MLOps
Frameworks	FastAPI · Django · Flask · Spring Boot · Angular 19 · Next.js · React · Flutter · Socket.IO · SQLAlchemy
Cloud/DevOps	AWS (Lambda/CDK/S3/Cognito) · Docker · Kubernetes · Helm · Jenkins · Git · CI/CD · SonarQube · Gradle
Protocols/Tools	MCP · A2A · gRPC · Protobuf · MQTT/mTLS · JWT · REST · WebSocket · VSS · CAN/DBC · Playwright · Postman · Jira

About me

Lead Engineer - Gen AI at Bosch Global Software Technologies with 6+ years of experience spanning Generative AI, LLM fine-tuning, functional safety (FuSA) AI, and cloud-native platform development. Led cross-functional teams building AI-powered products including RAG platforms, HARA (Hazard Analysis and Risk Assessment) AI for safety-critical automotive systems, voice-enabled developer tools, and multi-tenant SaaS infrastructure. Patent co-author, hackathon winner (Bosch AWS 2025, AppsForBharat 2025), and Bharat Mobility Expo 2025 presenter. Passionate about building complex safety-relevant AI systems that create meaningful impact in decision-making for functional requirements.

Work Experience

Bosch Global Software Technologies

Bangalore, India

Lead Engineer - Gen AI

Mar. 2023 - Current

- Won Bosch AWS Hackathon 2025 for ”Project IQ,” an AI-powered training & knowledge management system with 90% accuracy in extracting competencies from SRS documents.
- Won AppsForBharat 2025 hackathon for developing impactful solutions for India.
- Co-authored patent on detunneling data from ethernet frames using application-specific unique IDs (Patent No: 202541072960).
- Represented Bosch at Bharat Mobility Expo 2025, showcasing a Vehicle Assistant AI that generates SDV apps deployable on HMI and HPCs.
- Leading a cross-functional team of 8+ engineers across Gen AI, automotive middleware, and cloud platform initiatives at Bosch.
- Driving technical architecture decisions and conducting code reviews to ensure quality, scalability, and adherence to best practices.
- Mentoring junior developers on Gen AI, LLM fine-tuning, and cloud-native development, accelerating team ramp-up time.
- Coordinating with product owners, stakeholders, and global teams to translate business requirements into technical roadmaps and sprint deliverables.
- Establishing engineering standards for CI/CD pipelines, containerized deployments, and documentation across multiple SDV projects.
- Built HARA AI, a safety-relevant AI system for Hazard Analysis and Risk Assessment (HARA) that automates identification of safety goals and ASIL classifications per ISO 26262.
- Developed FuSA AI engine that analyzes functional requirements to generate safety-critical decision recommendations, reducing manual safety assessment effort.
- Engineered AI-driven analysis pipeline that extracts hazardous events from system-level functional requirements and maps them to appropriate safety mechanisms.
- Achieved 90% accuracy in extracting technical competencies from complex SRS documents, estimating €25,000+ in annual cost savings.
- Reduced manual training planning time by 80%, transforming a multi-hour task into a 3-second automated process.
- Engineered a semantic mapping engine for a 500+ skill taxonomy, solving inconsistent technical terminology across global teams.
- Architected BRICK, an advanced document analysis platform featuring PathRAG and MinerU for high-precision retrieval from unstructured PDFs with structural preservation.
- Developed a high-performance layout recognition engine using MinerU that identifies formulas, tables, and images with structural preservation.
- Implemented real-time annotation overlay for browser-based PDF viewing, enabling immediate visual verification of AI detections.
- Optimized a background processing queue for concurrent multi-format (PDF, DOCX) document analysis without blocking the UI.
- End-to-end architected HireStream, a recruitment OS matching associate skills against JDs using zero-shot learning, reducing manual screening by 80%.
- Developed Role-Based Dashboards for Talent Managers, Delivery Managers, and Associates with secure access to sensitive performance data.
- Automated extraction of skill levels (L0-L5) from legacy resumes and implemented an Intelligent Development Roadmap engine for personalized training.
- Engineered semantic matching logic that handles ”hidden skills” implied by experience, improving match quality over keyword search.
- Performed Parameter-Efficient Fine-Tuning (PEFT) on Qwen 7B+ models using QLoRA/LoRA with 4-bit quantization on limited hardware.
- Integrated Optuna for automated hyper-parameter tuning, achieving optimal training efficiency across multiple model trials.
- Curated and processed multi-million token datasets (jsonl) for specialized domain alignment in automotive software engineering.
- Developed SDX Assistant, a voice-enabled developer tool with STT/TTS and multi-turn conversation for natural language code generation.
- Integrated hands-free STT/TTS with custom wake-word support, enabling code generation while interacting with in-vehicle hardware.

- Designed a ”Check-and-Confirm” state machine that extracts VSS datapoints and provides pseudocode for validation before final code generation.
- Built SDX-A2L Signal Monitor with Two-Stage AI Search (20x faster) for parsing 100MB+ A2L files and real-time ECU memory monitoring via WebSocket.
- Engineered a C-Code generation engine that automatically translates monitored signals into embedded-ready structures.
- Solved the ”needle in a haystack” problem for ECU signal discovery using LLM-based semantic refinement of automotive terminology.
- Developed CAN-to-VSS protocol mapper translating low-level CAN bus (DBC) data into cloud-native VSS standards for MG Comet EV.
- Built a GUI-based mapper for engineers to visually define relationships between raw hex data and meaningful vehicle signals.
- Automated JSON-based mapping file generation, reducing manual protocol alignment time for new vehicle models.
- Optimized KUKSA.val databroker in C++/Rust for high-throughput, low-latency signal orchestration with secure gRPC/Protobuf communication.
- Engineered JWT-based authorization to restrict access to sensitive vehicle controls (HVAC, Windows) at the signal level.
- Reduced signal-to-app latency by optimizing internal data structures and memory management in the Databroker.
- Implemented enterprise-grade secure V2C telemetry agent using Mutual TLS (mTLS) with DigiCert certificate management.
- Developed high-reliability telemetry pipeline with local data persistence during connectivity outages and ordered signal delivery on reconnect.
- Hardened the agent against injection attacks by implementing strict VSS-based data sanitization protocols.
- Architected a Serverless Multi-Tenant SaaS platform using AWS CDK, reducing new vehicle project onboarding time by 90%.
- Built centralized Tenant Management Service controlling access, billing, and resource allocation across hundreds of AWS accounts.
- Optimized cloud infrastructure costs through auto-scaling Lambda functions and intelligent S3 storage tiering.
- Developed high-compliance backend microservices for Smart Connected EV platform using Java 17/Spring Boot with SonarQube/CycloneDX security.
- Architected a high-precision EV Trip Planning engine accounting for battery SoC, elevation, and charging station availability.
- Containerized and orchestrated global services using Docker and Kubernetes (Helm), ensuring 99.9% uptime for vehicle monitoring.

HashedIn by Deloitte

Bangalore, India

SDE - II

Aug. 2021 - Mar. 2023

- Built cloud-native content normalization platform for Thomson Reuters legal domain using AWS serverless architecture.
- Developed serverless applications using AWS Lambda, API Gateway, RDS, DynamoDB, AWS OCR, CodePipelines, and CloudFormation.
- Developed annotation component in Angular that reduced document content processing time from 7 days to 4 hours.
- Architected AWS serverless cloud infrastructure with relational and non-relational database integration.
- 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 and migrated database from Kinvey to Django models.
- Implemented FCM push notifications and Celery-based async task processing in Python.
- Created and maintained AWS instances for daily review extractions with NLP scoring using spaCy.
- Built Content Management service interface handling 1000+ client devices in WSO2.
- Redesigned 70% of UI controlling user flow architecture, improving user experience.
- Wrote automation test scripts in Python using Selenium, reducing manual testing effort by 35%.
- Collaborated on all stages of SDLC, from requirement gathering to production releases.
- Translated user requirements into project designs, implementation plans, and design mockups.

Zapcom Solutions

Bangalore, India

Full Stack Intern

Jan. 2019 - Jan. 2020

- Redesigned Zapcom portal with responsive landing pages in HTML/CSS, improving navigation and cross-browser compatibility.
- Built scene classification model using Places365 dataset for image recognition.
- Developed web scraping pipeline in Java (Selenium) and Python (Scrapy) to extract 1000+ reviews from 540+ client hotels.
- Integrated Facebook Developer API to automate review extraction for 100+ client hotels.

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	**Winner**, AppsForBharat 2025 -- Developing impactful solutions for India	*National*
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

Passion Projects

JobsChange.com -- AI Career Acceleration Platform

India

Co-Founder

2025 - Current

- Built an AI-powered career platform that tailors resumes to job descriptions, generates cover letters, and provides mock interview preparation using OpenAI GPT-5.
- Engineered full-stack application with Next.js 15, React 19, TypeScript, and Firebase, featuring real-time AI resume analysis and ATS compatibility scoring.
- Developed Chrome extension supporting 50+ ATS platforms (Workday, Greenhouse, Lever, Taleo) for one-click job application autofill.
- Implemented AI mock interview system with voice-based Q&A using Vapi, providing real-time feedback and performance scoring.
- Built integrated coding practice module with AI-generated problems, automated test case validation, and difficulty progression.
- Designed credit-based monetization system with Stripe payment integration and tiered subscription plans.

MakeDemos.com -- Professional Screen Recording & Demo Software

India

Co-Founder

2025 - Current

- Built a desktop screen recording application for creating polished product demos with auto-zoom, 50+ visual effects, and 4K export capabilities.
- Engineered cross-platform desktop app using Electron 30, React 18, and PixiJS 8 with GPU-accelerated rendering pipeline via Web Codecs API.
- Implemented intelligent auto-zoom that tracks cursor movements and click events, automatically framing key actions during recordings.
- Developed real-time video effects engine with PixiJS shaders supporting transitions, annotations, blur, and spotlight effects on timeline.
- Built multi-track timeline editor supporting screen, camera, and audio tracks with frame-accurate trimming and export.
- Designed state management architecture using Zustand for real-time recording controls and Firebase for user authentication and licensing.