Lahari Karrotu

SOFTWARE ENGINEER | BACKEND & AI SYSTEMS

San Jose, CA | laharikarrotu24@gmail.com | linkedin.com/in/lahari-karrotu | laharikarrotuportfolio.site

SUMMARY

Software Engineer focused on backend systems and production AI applications. Experienced in building distributed services, LLM-powered systems, and scalable full-stack integrations under real-world constraints (latency, reliability, observability). Passionate about creating maintainable, high-performance solutions that bridge AI models with production-ready systems.

PROFESSIONAL EXPERIENCE

Arkatech Solutions

Software Engineer (Full Stack AI)

May 2025 – Present

- Designed and deployed backend LLM services using Python and Ray to support distributed inference and concurrent request handling.
- Built multimodal AI pipelines combining text and image embeddings, stored and retrieved using Milvus and Pinecone for low-latency semantic search.
- Developed REST APIs with Flask to expose AI capabilities to frontend and product teams.
- Implemented structured logging, monitoring, and fault-tolerant workflows to improve reliability of production AI systems.
- Optimized inference performance through batching and orchestration strategies in distributed environments.

Anguliyam AI Solutions

Software Engineer Intern (AI & Full-Stack)

Jun 2024 – May 2025

- Developed LLM-driven voice and text workflows using Python and Flask for real-time interaction.
- Built agent orchestration frameworks to manage multi-step API chaining and tool execution.
- Optimized embedding preprocessing and retrieval pipelines to improve efficiency and response latency.
- Integrated backend AI services into end-to-end systems supporting production use cases.

Cognizant Technology Solutions

Program Analyst Intern

Jan 2022 – Aug 2022

- Supported large-scale ML workloads by building ETL and data preprocessing pipelines in Python.
- Improved model throughput by optimizing data flows and caching strategies.
- Collaborated on NLP and structured data processing modules for downstream ML systems.

EPAM Systems
Software Engineer Intern
Dec 2020 – Mar 2021

- Developed automation scripts to streamline data and ML workflows.
- Assisted in feature engineering and batch processing for analytics pipelines.
- Improved operational efficiency of reporting and data handling systems.

PROJECTS

ScanX — Real-Time Scan-to-Action AI System

Python, FastAPI, Next.js, React Native • GitHub

- Enabled automated execution of tasks across complex interfaces using a real-time AI agent that interprets UI with LLMs.
- Built multimodal vision + planning pipelines and scalable backend APIs with FastAPI.
- Integrated frontend (Next.js/React Native) for seamless user interactions, improving efficiency and accuracy.

Blinds & Boundaries — Vision-Based Perception and Decision System

React, TypeScript, FastAPI, Python, Azure • Github

- Developed a virtual try-on system for window blinds using computer vision and 3D rendering.
- Implemented real-time backend inference pipelines with FastAPI, and integrated a responsive frontend.
- Deployed on Azure cloud for low-latency and scalable performance, enabling realistic visualizations for users.

TECHNICAL SKILLS

Programming & Scripting: Python, Java, JavaScript, Go, C++, Erlang
Backend / APIs: Flask, FastAPI, Express.js, Node.js, Django, Golang
Frontend / Full Stack: React, React Native, HTML, CSS, TypeScript, Redux
Databases & Storage: MySQL, MongoDB, SQLite, Milvus, Pinecone, DB2
Cloud & DevOps: AWS, Azure, Docker, Ray, LangChain, Streamlit
AI & ML: LLMs, NLP, Multimodal AI pipelines, Embeddings, Fine-tuning
Tools & Monitoring: Git/GitHub, LangSmith, Langfuse, Jira, Trello

EDUCATION

Master of Science in Computer Science, Florida Institute of Technology, 2024 — GPA: 3.5/4.0
Bachelor of Technology in Computer Science, KL University, 2022 — GPA: 8.9/10.0

CERTIFICATIONS

AWS Certified Solutions Architect – Associate
Cisco CCNA: Switching, Routing, and Wireless Essentials