Lahari Karrotu

San Jose,CA | +1 321-234-6914 | laharikarrotu24@gmail.com
Laharikarrotuportfolio.site | www.linkedin.com/in/laharikarrotu | github.com/laharikarrotu

Summary

AI/ML & Full-Stack Engineer experienced in building scalable distributed systems, real-time inference pipelines, and production-grade AI applications. Skilled in LLM-based agent architectures, microservices, cloud-native engineering, and end-to-end platform development. I design reliable, high-performance systems that combine strong software foundations with modern AI capabilities across the full stack.

Core Competencies: LLMs • Distributed Systems • Full-Stack Engineering • Cloud Architecture • MLOps • Microservices • System Design • Data Engineering

Education

Florida Institute of Technology **Aug 2022 - May 2024**
Master of Science, Computer Science **GPA:** 3.6/4
- **Coursework:** Artificial Intelligence, Cloud Computing, Computer Networks, Speech Recognition, Analysis of Algorithms, Computer Information Systems

KL University **Aug 2018 - Jun 2022**
Bachelor of Science, Computer Science **GPA:** 3.6/4
- **Coursework:** Data Structures & Algorithms, Operating Systems, DBMS, Data Science, Computer Architecture, Software Engineering

Technical Skills

- **Programming & Core CS :** Python, SQL, PySpark, JavaScript, TypeScript, Java, Data Structures & Algorithms, Problem Solving, Debugging & Optimization
- **AI / Machine Learning :** TensorFlow, PyTorch, Hugging Face, LangChain, RAG Pipelines, Embedding Models, Computer Vision, NLP, Model Training & Evaluation, Real-Time Inference Systems
- **LLM / Agent Engineering :** Vision-Language Models (GPT-4o, Claude, Gemini), Tool-Use Agents, Planning & Execution Modules, Autonomous Workflow Agents, Prompt Engineering
- **System Design & Architecture :** Microservices Architecture, Distributed Systems, Event-Driven Design, API Architecture, Caching & Load Balancing, Serverless Patterns, Concurrency & Asynchronous Execution, Fault-Tolerant Pipelines
- **Cloud Engineering :** Azure (ADF, Synapse, App Service, Cognitive Services), AWS (SageMaker, Lambda, Glue, Redshift), GCP (Vertex AI, BigQuery, Firebase), Cloud Storage, Auto-Scaling, Load Balancers, IAM & Security
- **MLOps / DevOps :** MLflow, Docker, Kubernetes, Terraform, GitHub Actions, CI/CD Pipelines, Model Deployment, Model Registry • Monitoring & Observability
- **Data Engineering :** Apache Spark, Databricks, Kafka, Airflow, Snowflake, Delta Lake, ETL/ELT, Data Modeling, Batch & Stream Processing, Data Quality & Validation
- **Full-Stack Development :** React, Next.js, TypeScript, FastAPI, Node.js, REST APIs, Component Architecture
- **Databases :** PostgreSQL, MongoDB, MySQL, Redis, Elasticsearch
- **Tools & Visualization :** Git, GitHub, Linux, Postman, Grafana, Power BI, Tableau

Experience

Software Engineer – AI & Data Systems
Arkatech Solutions **May 2025 – Present | Remote**

- Designed and implemented end-to-end AI and data systems using microservices, distributed storage, and event-driven architecture.
- Built LLM-powered automation and personalization pipelines using embeddings, retrieval layers, and modular orchestration workflows.
- Developed full-stack features with React/Next.js (frontend) and FastAPI/Node.js (backend) supporting real-time AI interactions.
- Created real-time inference APIs with secure gateways, autoscaling, and cloud-native deployment patterns on Azure.
- Implemented MLOps pipelines using MLflow, Docker, Terraform, and CI/CD (GitHub Actions) to standardize training, deployment, and monitoring.

Software Engineer Intern – AI & Full-Stack Systems
Anguliyam AI Solutions **Jun 2024 – Apr 2025 | Remote**

- Developed SmartBuy AI Assistant and Navigating Assistant, enabling product discovery and guided user flows using LLMs, embeddings, and vector-based retrieval.
- Built a voice-enabled assistant with speech-to-text and text-to-speech integration for real-time conversational interactions.
- Implemented full-stack features using React/Next.js (UI) and FastAPI/Node.js (backend) to support AI actions, routing, and user session handling.
- Designed prompt orchestration and multi-tool agent logic, improving accuracy and personalization across workflows.

- Optimized backend performance with async processing, caching strategies, and modular microservices, improving latency and scalability.

Data Engineer Intern – Cloud & Big Data

Cognizant Technology Solutions Jan 2022 – Aug 2022 | Bengaluru, India

- Built scalable ETL/ELT pipelines using AWS Glue, Spark, and Redshift for enterprise analytics workloads.
- Developed real-time streaming ingestion using Kafka and Lambda to automate data freshness.
- Optimized distributed Spark workloads with partitioning, caching, and broadcast joins, reducing compute time and cost.
- Supported data modeling, transformation logic, validation rules, and reusable components across cross-functional teams.

Data Engineering Intern – Analytics Automation

EPAM Systems Dec 2020 – Mar 2021 | Hyderabad, India

- Automated batch and streaming workflows using PySpark on EMR with Airflow DAGs, improving pipeline stability.
- Performed data validation, anomaly detection, and quality scoring, enhancing dataset accuracy for analytics and ML use cases.
- Built dashboards with Tableau and Power BI to support KPI visibility and operational reporting.
- Developed automated reporting pipelines, reducing recurring manual workload for business teams.

Projects

ScanToActionAI – Real-Time Scan-to-Action Agent

Python • FastAPI • OpenCV • Agent Looping • Orchestration

- Developed a real-time UI understanding agent that captures screens, detects interface components using OpenCV, and triggers automated actions through a Python agent loop.
- Designed a backend service to manage frame ingestion, action execution, state tracking, and multi-step workflows.
- Implemented modular pipelines supporting OCR, element detection, and event routing for extensible agent capabilities.
- Optimized frame processing and inference latency through async loops, batched operations, and caching strategies.

SmartBuy AI – AI Navigation & Recommendation Platform

Next.js • FastAPI • PostgreSQL • Vector Search • LLM Orchestration

- Built an end-to-end AI shopping assistant enabling navigation, product search, and personalized recommendations using embeddings and vector similarity search.
- Engineered backend services with FastAPI + PostgreSQL, supporting secure routing, session handling, user context, and ingestion pipelines.
- Integrated Next.js server actions for SSR/ISR rendering, improving UI responsiveness and SEO performance.
- Implemented non-functional improvements including API rate-limiting, caching, async execution, and database query optimization.

Voice AI Assistant – Real-Time Conversational Agent

Speech-to-Text • TTS • FastAPI Streaming • WebSockets • Cloud Deployment

- Developed a voice-enabled AI assistant with speech recognition, TTS synthesis, and multi-turn conversational orchestration.
- Architected a backend using FastAPI streaming endpoints + Web-sockets to support low-latency real-time conversation flows.
- Integrated LLM-driven agent actions enabling navigation, contextual decision-making, and guided multi-step tasks.
- Deployed cloud-hosted inference endpoints with autoscaling, observability, and async workers to maintain stability at scale.

Predictive Maintenance System – Manufacturing/IoT Pipeline

Apache Spark • Snowflake • Azure Data Lake • TensorFlow • REST API

- Built a distributed data pipeline for ingesting, cleaning, and analyzing high-frequency sensor streams using Spark and Delta Lake.
- Trained LSTM / XGBoost models for early failure prediction and deployed inference endpoints for streaming alert generation.
- Designed non-functional requirements including fault tolerance, schema enforcement, and time-series aggregation optimizations.
- Delivered dashboards and REST APIs exposing live operational insights to manufacturing stakeholders.

Blinds & Boundaries – Full-Stack Visualization System

React • FastAPI • Computer Vision • Azure Storage • Microservices

- Created a virtual try-on system that overlays blinds/curtains onto room images using CV-based room analysis, edge detection, and perspective transforms.
- Built back-end microservices exposing CV inference, product catalog APIs, and image transformation pipelines.
- Implemented React-based dynamic rendering, caching, optimized asset loading, and responsive UI flows.
- Deployed the system using containerized services with cloud storage integrations for scalable asset retrieval.

Certifications

Machine Learning Specialization – Andrew Ng (Coursera)
AWS Certified Solutions Architect – Associate (2025)
Cisco Certified Network Associate (CCNA)
Oracle Database SQL Certified Associate