LALITH SAGAR KAMBALA

KS, United States (Open to Relocate) | (913)-488-0303 | lalithsagarkambala@gmail.com | LinkedIn | Portfolio| Github

SUMMARY

Software Engineer with a founder mindset who thrives in 0→1 environments, owning problems end-to-end and turning fuzzy ideas into reliable products. I care deeply about user experience, move quickly with high standards, and balance experimentation with pragmatic, production-ready engineering.

EDUCATION

Master of Science: Computer Science — Aug 2023 to May 2025
University of Central Missouri - Lee's Summit, MO – **GPA**: 3.55/4.00

EXPERIENCE

Founding Software Engineer — Sep 2025 to Present
Todaiyo.ai – United States

- Owned and scaled backend systems end-to-end, from architecture and API design through deployment and monitoring, while collaborating closely on UI/UX flows to ensure fast, intuitive, user-facing experiences.
- Built asynchronous, event-driven workflows with queues and workers to support concurrent AI-powered requests and real-time user interactions under production traffic.
- Designed resilient backend services with reliability and uptime as core goals, proactively debugging, triaging, and resolving production incidents.
- Led full-stack development of a production SaaS product using React, Node.js, and Supabase, integrating AI APIs, reducing OpenAI costs by 25%, implementing image-generation pipelines, and instrumenting analytics with Google Analytics and Google Tag Manager to improve engagement metrics.

AI Full Stack Engineer *(Part-time)* — Sep 2025 to Present
Consumer Genie – United States

- Architected and built an agentic AI-powered shopping platform that reduced cross-site product research time by 40%, leveraging LangGraph to model multi-agent reasoning flows for search, summarization, ranking, and comparison.
- Designed a distributed, event-driven orchestration system with asynchronous task execution to coordinate autonomous AI agents for scalable, fault-tolerant handling of complex multi-step user queries.
- Implemented a semantic search engine using vector embeddings with MongoDB-based index storage, integrating DuckDuckGo, Serper, and Tavily APIs for real-time discovery, review aggregation, and intelligent product recommendations.
- Building the real-time, low-latency AI pipelines on AWS using Docker and CI/CD automation, applying LLM guardrails and cost-control mechanisms to optimize API spend and ensure concurrency under load.

Full Stack Software Engineer — Jul 2025 to Oct 2025
IEJL – United States

- Engineering full-stack web applications and internal tools using Next.js, TypeScript, and Tailwind CSS, improving UI load times by 30% while delivering scalable backend APIs that handle concurrent requests, real-time updates, and Supabase-powered PostgreSQL data synchronization, boosting data sync efficiency by 25%.
- Optimized PostgreSQL schemas, indexes, and queries (via Supabase) to improve performance, reliability, and consistency of high-throughput production workloads.
- Partnered with product managers, UI/UX designers, and cross-functional teams in Agile sprints, accelerating feature delivery and reducing iteration cycles by 35% through close collaboration and rapid feedback loops.
- Implemented Deployed containerized services on AWS with Docker and CI/CD pipelines, improving system stability, enabling faster and safer production releases, and achieving 40% faster deployment cycles and 30% better system reliability.

Web Development Intern — Sep 2022 to Nov 2022
Coincent Technologies – India

- Created a user-friendly Job Portal Management System with optimized MySQL queries and modular front-end components, reducing job posting and record processing time by 30% and enabling smoother applicant navigation.
- Enhanced application reliability by validating mobile-first user flows and cross-browser compatibility, while supporting Java **(**Spring Boot**)** backend APIs and monitoring logs and errors, improving accessibility for 95% of users and reducing bounce rates by 20%.
- Debugged and troubleshot system issues using automated error monitoring and rapid resolution workflows, maintaining 99% uptime and ensuring uninterrupted user access during high-traffic periods.

ACHIEVEMENTS & CERTIFICATIONS

- Oracle Cloud Infrastructure 2024 Generative AI Professional (1Z0-1127-24)
- Introduction to Large Language Models (LLM) by Google Cloud
- Accenture North America Data Science and Visualization Job Simulation on Forage
- Smart India Hackathon Winner for Blockchain and decentralization of tax collections

COMPETENCIES

- Programming languages: Python, Java(SpringBoot), C#, SQL, HTML, Tailwind CSS, JavaScript, TypeScript, Rest APIs
- Skills: Data Structures & Algorithms, UI/UX designing, Data Modelling, Mining, Data Visualization, AI, LangChain, Machine Learning
- Tech: VScode, Jupyter, Colab, Linux, Git, Jira, Selenium, Junit Testing, GenAI, Vector database, Spring Boot, CursorAI, Stripe, Material UI
- Frameworks: ReactJS, Flask, TensorFlow, scikit-learn, Pandas, Keras, Matplotlib, Seaborn, Tableau, MongoDB
- Coursework: Relational Database, Cloud Computing, Operating Systems, SDLC & Agile Methodologies (Scrum), Adv Software Engineering
- Soft Skills: Teamwork and collaboration, Communication skills

PROJECTS

AI- Debugging Agent using ollama – Pinecone **Mar 2026**

- Developed an end-to-end AI Debugging Agent using FastAPI and Next.js (TypeScript), capable of ingesting GitHub and ZIP codebases, chunking and embedding source files, and answering natural-language debugging queries with precise file/line references.
- Designed a modular Retrieval-Augmented Generation (RAG) pipeline leveraging open-source models (SentenceTransformers, CodeLlama via Ollama), featuring retrieval orchestration, structured JSON outputs, query history tracking via SQLite, and optional static analysis integration (pylint/eslint) for enhanced debugging context.
- Implemented production-grade engineering practices, including detailed request and pipeline logging, robust error handling with LLM/Pinecone fallback strategies, namespace-based multi-repo querying with a saved repo dropdown, and a privacy-first local vector store mode to maintain repository data on-device.

Genie Shop (Consumer Genie) **Sep 2025**

- Developed a product search system using LangGraph state machine with intent-based routing that processes research and specific product queries through separate pipelines, streaming real-time results via Streamlit.
- Built web scraping pipeline using Google Gemini for intent classification and product extraction, integrating DuckDuckGo/Serper API to search across major retailers with URL scoring and filtering mechanisms.
- Created location-aware cross-platform retrieval system with category-based site selection across USA/Canada markets, generating automated fallback search URLs and implementing multi-model retry strategies that ensure 95%+ query completion rate despite API failures.

Article Semantic Search using LLM's **Aug 2025**

- Designed a financial news research platform leveraging LangChain’s Unstructured URL Loader and OpenAI embeddings for precise content indexing and retrieval.
- Deployed FAISS-based semantic search with embedding vector storage and built an interactive Streamlit app allowing users to process URLs and obtain contextually accurate answers from a large language model.

Gemini API Clone **May 2025 - Jun 2025**

- Crafted a responsive Gemini Clone web application using HTML, CSS, JavaScript, JSX, and React.js, focusing on modular components and seamless user interaction.
- Integrated RESTful APIs and asynchronous JavaScript to enable instant data updates and build scalable, production-ready interfaces

Brain Tumor Detection Using Deep Learning **Jan 2023 - Apr 2023**

- Constructed a CNN model with Conv2D, Batch Normalization, MaxPooling, Dropout, and Dense layers using ReLU activation to classify MRI scans with 94% accuracy, enhancing feature extraction and reducing overfitting by 15%.
- Automated early tumor detection workflows, cutting diagnosis time by up to 40% and improving patient care through scalable, real-time image analysis.

Blockchain project on Property Management System **Aug 2022 - Dec 2022**

- Built **Block Wizz**, a decentralized blockchain-based tax payment solution that processed 500+ transactions with 100% transparency, ensuring tamper-proof records and full auditability across the network.
- Incorporated SHA-256 hashing to securely chain payment histories, reducing fraudulent receipts and duplicate payments by 95% while enabling real-time, verifiable transaction tracking.