Lakshmi Swathi Sreedhar

AI Engineer | Product-Driven & Customer-Focused Builder | Generative AI Systems Developer

Lansing, MI · +1-3135292810 · swathisl@umich.edu

LinkedIn | GitHub | Blog |Portfolio

PROFILE SUMMARY

AI Engineer with deep experience building production-grade Generative AI systems, including LLM orchestration, retrieval-augmented generation (RAG), and multi-agent workflows across both enterprise and early-stage startup environments. Specializes in designing scalable, reliable backend AI architectures that combine model development, API services, and orchestration layers to support complex reasoning, validation, and automation at scale. Operates with a forward-deployed, consulting-style approach, working closely with product, data, and engineering stakeholders to translate ambiguous requirements into robust AI systems that improve decision accuracy, strengthen data quality, and accelerate time-to-insight. I bring a builder's mindset with strong system-level thinking, emphasizing reusability, fault tolerance, observability, and maintainability in AI-driven platforms, and am motivated to lead end-to-end architectures and 0–1 initiative shaping production AI solutions.

SKILLS

Programming & Data: Python, SQL (PostgreSQL, MySQL, Databricks), PySpark, Pandas, NumPy, Bash
AI & ML Systems: Machine learning pipelines, recommendation systems, A/B testing, explainable AI (XAI)
Generative AI & LLM Systems: Large Language Models (LLMs), deep learning (neural networks), multi-agentic workflows, prompt engineering, multimodal systems
LLM Orchestration & Tooling: Retrieval-Augmented Generation (RAG), LangChain, LangGraph, DSPy, Model Context Protocol (MCP), MCP servers, FAISS, Pinecone, PyDantic.
Data Engineering & Governance: ETL, Data Governance, Data Quality & Metadata Management.
AI Copilot & Developer Workflows: Cursor, Claude Code
Backend, Data & MLOps: FastAPI, Flask, async APIs, Apache Airflow, Docker, MLflow, CI/CD
Cloud & No-Code / Low-Code Platforms: Google Cloud Platform (Vertex AI, BigQuery), Microsoft Azure (Azure OpenAI, Synapse), Ikigai, Google Teachable Machine
Collaboration & Delivery: Git/GitHub, Agile/Scrum, Jira, Confluence
Expanding Tech Stack: React.js, Node.js, Next.js, TypeScript

PROFESSIONAL EXPERIENCE

CHAINSYS CORPORATION **09/2023 – 05/2025** Grand Ledge, MI

AI/ML -Software Engineer

- Joined ChainSys as one of the earliest engineers in the newly formed AI/ML team, helping define architecture standards, reusable platform patterns, tooling, and delivery practices for enterprise AI initiatives within a 500–1000 employee organization.
- Led customer pilots and AI workflow PoCs, translating complex client requirements into scalable, production-ready AI solutions that reduced manual workloads by 60–80% across healthcare and supply chain domains.
- Engineered production-grade backend AI services using async FastAPI and Python, building high-throughput, reliable orchestration layers for LLM reasoning, retrieval, and metadata-driven workflows.
- Architected an enterprise Text2SQL + RAG platform (LangChain, LangGraph, DSPy, FAISS, Pinecone) as a reusable internal product, improving SQL accuracy by 35% and reducing query turnaround time by 80%.
- Designed a multi-agent SQL reasoning architecture (Schema → Planner → Generator → Validator → Critic) with explicit state management and validation, reducing invalid SQL generation by 70% while delivering consistent, explainable outputs.
- Built MCP-style servers and tool schemas to standardize agent–tool interaction, enabling reusable, composable agents and consistent context handling across multiple AI workflows.
- Optimized semantic retrieval using hybrid FAISS + Pinecone indexing, improving grounding relevance and context matching by 30–40% for enterprise-scale datasets.
- Integrated multiple LLM families (GPT-4, GPT-5 previews, Claude, Mistral Codestral on Vertex AI) behind model-agnostic orchestration layers, ensuring flexibility, scalability, and controlled experimentation across client use cases.
- Used AI copilots (Cursor, Claude Code) for debugging, prompt iteration, agent behavior analysis, and rapid validation during development of complex orchestration and multi-agent workflows.
- Developed reusable automation agents (retrieval, validation, mapping, anomaly detection, summarization) packaged as platform components, accelerating customer onboarding by 25–30% and standardizing AI behavior across implementations.
- Enhanced ETL and data migration pipelines by embedding LLM-driven schema alignment, mapping inference, and quality validation, saving data engineering teams hundreds of manual hours per migration cycle.
- Built a BERT-based data deduplication and data quality engine integrated with ChainSys platform rules, improving dataset reliability by 40% and reducing downstream cleansing effort.
- Implemented monitoring and evaluation dashboards (Streamlit, internal tools) for model performance, query accuracy, and data-quality metrics, improving observability, reliability, and stakeholder trust.
- Containerized AI microservices with Docker, applied MLflow for experiment tracking and governance, and deployed pipelines across Azure ML and GCP Vertex AI, ensuring reproducible, scalable, and stable production lifecycles.
- Mentored junior engineers on agent architecture, orchestration patterns, retrieval optimization, evaluation frameworks, and system reliability, raising overall delivery quality across the AI/ML team.

PARAILLEL INC (EARLY-STAGE STARTUP) **03/2023 – 08/2023** Detroit, MI

Founding Machine Learning Developer Intern

- Joined Paraillel as one of the founding ML interns during its earliest startup phase, helping build the first versions of the company's personalization and content intelligence systems in a fast-paced 0→1 environment.
- Designed and implemented an AI-driven Recommendation Engine combining collaborative filtering with GAN-based generative augmentation to improve personalization and cold-start performance across multi-domain user datasets.
- Integrated transformer and BERT embeddings to enrich semantic understanding for ranking and similarity scoring, increasing recommendation precision and diversity by 25% while reducing sparse-data failures.
- Explored agentic-style workflow patterns, experimenting with modular components for ranking, re-ranking, enrichment, and metadata interpretation that formed the foundation for adaptive recommendation behaviours.
- Built multimodal input pipelines blending textual, behavioral, and categorical features to enable adaptive recommendation flows for e-commerce and media use cases.
- Developed explainability and evaluation dashboards using Streamlit and Plotly to visualize latent features, model reasoning, and user-level insights strengthening transparency for product stakeholders.
- Optimized retraining and deployment cycles using MLflow, Docker, and TensorFlow Recommenders, enabling reproducible experiments, versioning discipline, and smoother CI/CD integration.
- Collaborated closely with product and engineering teams to translate personalization goals into measurable AI outcomes aligned with user engagement KPIs and platform scalability requirements.
- Contributed to internal research on generative personalization, exploring hybrid architectures that combine LLMs, embeddings, and retrieval components for context-aware, conversational discovery experiences.
- Gained hands-on experience with rapid prototyping, ambiguity navigation, multi-role execution, and startup-style experimentation cycles.

APTECH INC **12/2020 – 04/2021** **BANGALORE, INDIA**

AI/ML Engineer Trainee

- Completed 400+ hours of intensive training in AI/ML fundamentals, software development, and product engineering concepts.
- Constructed machine learning pipelines for classification, clustering, regression, and forecasting using Scikit-learn.
- Developed RESTful APIs and interactive GUIs with Flask and PyQt, integrating ML models into user-facing applications.
- Applied statistical methods, linear algebra, and EDA techniques to analyse and optimize predictive models.
- Practiced agile workflows, version control with Git, and real-time debugging in a simulated production environment.

EDUCATION

MIT x PRO – MASSACHUSETTS INSTITUTE OF TECHNOLOGY **02/2025 – 06/2025** REMOTE
No-Code AI and Machine Learning: Building Data Science Solutions

UNIVERSITY OF MICHIGAN DEARBORN **09/2021-04/ 2023** Dearborn, MI
Masters in Artificial Intelligence; Minor in Machine Learning
- Non-Resident Graduate Scholarship Recipient

DAYANANDA SAGAR COLLEGE OF ENGINEERING **08/2016 – 08/ 2020** Bangalore, India
Bachelors in Aeronautical Engineering

PROJECTS / OPEN-SOURCE

- **AI-Powered Text2SQL & Agentic RAG System** – Python, GPT-4, LangChain, LangGraph, FAISS, DSPy, PyDantic, Streamlit, PostgreSQL, Oracle, Databricks
- **Clinical RAG Data Extraction Pipeline** – Python, Hugging Face, FAISS, Azure OpenAI, Pandas
- **Autonomous-Driving TurtleBot** – MATLAB, ROS, Lidar, Image Processing
- **AI-Powered UNSPSC Category Generator** – Python, GPT-3.5 API, Pandas, Excel
- **Semantic Segmentation for Autonomous Driving** – Python, PyTorch, TensorFlow, OpenCV, ROS

☞ *Full project details, write-ups, and code available at: My Projects*

CERTIFICATIONS

- No Code AI and Machine Learning: Building Data Science Solutions by MIT professional Education –MIT

☞ *More Skills and Certifications available at: Certifications*

PUBLICATIONS AND PRESENTATION

- ***Advancing Precision Medicine through Multimodal AI: Innovative Approaches to Diagnostics and Treatment*** – Proposes a multimodal AI framework combining medical imaging, clinical text, and wearable data to improve diagnostics and personalized treatments while ensuring data privacy with federated learning.
- ***Speak the Language of AI: Mastering Prompt Engineering for LLMs –*** Explores advanced prompt engineering techniques for LLMs, introducing automated tuning frameworks to improve performance across business, education, and healthcare.