

Kavinraja D

Chennai, India | Open to Remote / Bengaluru | kavinraja.d1612@gmail.com | +91-8148472763
github.com/d-kavinraja | linkedin.com/in/kavinraja-d

Professional Summary

AI Engineer with hands-on experience developing, evaluating, and deploying Generative AI solutions, RAG workflows, LLM applications, and production REST APIs. Proficient in Python, LLM integrations, prompt template engineering, vector databases (ChromaDB), and cloud platforms (AWS/Azure). Skilled in context retrieval optimization, solution testing, data preprocessing pipelines, containerization (Docker/CI-CD), and troubleshooting operational failure cases to build reliable enterprise AI capabilities.

Education

Saveetha Engineering College, Chennai, India
B.Tech in Artificial Intelligence and Machine Learning

August 2022 – May 2026
CGPA: 8.6 / 10

Technical Skills

Generative AI & LLM Engineering: Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Prompt Engineering & Evaluation, Vector Databases (ChromaDB), AI Workflow Orchestration, Context Embeddings, NLP

Machine Learning & Frameworks: PyTorch, TensorFlow, Scikit-Learn, Statistical Modeling, Model Training & Validation Metrics, Deep Learning (YOLO, CRNN, 3D CNN, BiLSTM), Computer Vision

Software Engineering & APIs: Python, REST APIs, FastAPI, JSON Data Contracts, PostgreSQL, SQL, Redis, Data Preprocessing & Ingestion Pipelines, Object-Oriented Programming (OOP)

MLOps, Observability & Operations: Docker, CI/CD (GitHub Actions), System & Log Monitoring, Latency & Error Troubleshooting, Automated Unit Testing, Root Cause Analysis (RCA)

Cloud Platform & Tools: AWS Certified AI Practitioner, AWS Certified Cloud Practitioner, Linux, Version Control (Git/GitHub)

Professional Experience

Cognizant Technology Solutions — Program Analyst Intern (Infrastructure & Cloud Support) *March 2026 – July 2026*

- Supported enterprise cloud and infrastructure environments across Linux, Windows Server, Azure, and VDI platforms, developing deep operational understanding of system and application-layer behavior.
- Analyzed system logs, monitoring metrics, and incident data to investigate operational failures, troubleshoot performance issues, and perform data-driven root cause analysis (RCA).
- Authored technical documentation, operational runbooks, and knowledge guides to support consistent issue resolution and maintainable engineering practices.

Sasken Technologies Ltd. — GenAI & Data Science Intern *June 2025 – July 2025*

- Built Python data preprocessing pipelines and optimized vector-based context retrieval (RAG) using ChromaDB for Generative AI workflows, improving context retrieval performance by 40%.
- Constructed testing workflows across 10 GB+ code repositories using statistical analysis and accuracy benchmarks, evaluating retrieval output reliability at 92%.
- Authored technical specifications, REST API documentation, implementation workflows, and evaluation reports in collaboration with cross-functional engineering teams.

Projects

Revora — Open-Source AI Code Analysis Platform [\[GitHub\]](#) | *May 2026 – August 2026*

- Engineered a repository-aware Generative AI platform using Python, FastAPI, and LLM APIs to execute multi-step analysis spanning context retrieval, prompt construction, LLM orchestration, and output verification.
- Designed PostgreSQL database schemas for metadata, review history, and audit tracking, integrating Redis for asynchronous task processing and Docker for containerized development workflows.
- Implemented RESTful application services with structured JSON payloads, prompt validation, error handling, and automated GitHub Actions CI/CD workflows to ensure solution quality and reliability.

Intelligent Examination Submission Framework for LMS [\[GitHub\]](#) | *November 2025 – April 2026*

- Developed an end-to-end document intelligence pipeline using YOLO and CRNN deep learning models to extract structured information from unstructured documents, achieving 98.5% precision.
- Built FastAPI REST endpoints and PostgreSQL database layers supporting automated batch data ingestion and structured JSON transformation workflows, reducing manual processing effort by 70%.
- Integrated input validation, unit testing suites, and CI/CD pipelines to monitor processing failure cases.

MouthMap — 3D CNN + BiLSTM Lip Reading System [\[GitHub\]](#) | *March 2025 – October 2025*

- Developed an end-to-end silent lip-reading system using a 3D CNN + BiLSTM architecture, implementing video preprocessing, frame extraction, and CTC sequence modeling with TensorFlow and OpenCV.
- Evaluated 1,247 video samples across configurable neural network training and inference pipelines, achieving 89.3% prediction accuracy.
- Built model execution workflows and deployed an interactive Streamlit application to evaluate model output quality on real-time video inputs.

Certifications

- AWS Certified AI Practitioner** — Amazon Web Services [\[Credential\]](#)
- AWS Certified Cloud Practitioner** — Amazon Web Services [\[Credential\]](#)