

Vishnu Sai Varshith Kaki

(352) 219-1474 | vishnusaivarshith@gmail.com | linkedin.com/in/varshith-kaki | github.com/varshith-k | varshithkaki.work

EDUCATION

University of Florida

Master of Science in Computer Science — GPA: 4.0/4.0

Aug 2025 – Dec 2026

Gainesville, FL

Indian Institute of Information Technology, Sri City

Bachelor of Technology in Computer Science — CGPA: 9.32/10

Aug 2019 – May 2023

Sri City, India

EXPERIENCE

Application Developer

UF College of Medicine, Dept. of Health Outcomes & Biomedical Informatics

2026 – Present

Remote

- Healthcare AI research: data cleaning/extraction and NLP/ML model development for clinical decision support.

Agentic AI & Platform Engineering Intern

UF Information Technology (UFIT)

Apr 2026 – Sep 2026

Gainesville, FL

- Building agentic AI workflows and automation for the NaviGator AI platform (n8n, LangChain, OpenAI APIs); created reusable prompt libraries and GenAI evaluation pipelines.
- Fine-tuning AI responses on domain-specific datasets and running adversarial and regression testing to improve relevance, accuracy, and reliability.

AI/ML Intern

Ensora Health

May 2026 – Aug 2026

Orlando, FL

- Architected end-to-end RAG systems with vector databases for context-aware healthcare document retrieval.
- Built an AI-powered RCM Highlights Agent (LLMs + workflow orchestration) extracting insights from Revenue Cycle Management data and generating automated stakeholder-ready reports.
- Benchmarked LLM frameworks (LangChain, LlamaIndex) on accuracy, latency, scalability, and cost.

Software Engineer

Couture AI

Jul 2022 – Aug 2025

Bengaluru, India

- Developed production LLM-powered microservices (summarization, classification, structured extraction) with prompt templates, output schema validation, and logging for reliable deployments.
- Built LangChain RAG with FAISS vector search as a Dockerized inference API (Prometheus/Grafana monitoring); engineered LangGraph multi-agent workflows with guardrails and fallbacks.
- Designed large-scale ETL pipelines (Azure Data Factory, Databricks) for pre-processing high-volume datasets; optimized runtimes via modularization and parallelism.
- Revamped the Python backend of a global banking system (encryption, parallelization, modularization), cutting API latency 80%+; built Django/Flask REST APIs serving millions of daily requests.

PROJECTS

RetailBrain | *Python, FastAPI, Kafka (Redpanda), PostgreSQL, LangChain, LangGraph, Docker*

- Built an event-driven retail monitoring platform streaming live sales events; LangChain/LangGraph AI layer answers business questions in natural language.
- Containerized the full stack with Docker Compose; added Prometheus/Grafana observability for latency and anomaly rates.

CryptoPulse AI | *React, TypeScript, FastAPI, Kafka, Spark, PostgreSQL*

- Engineered a real-time crypto analytics platform computing technical indicators over streaming market data; grounded AI agent explains market movements and flags anomalies.
- Production-minded design: JWT auth, Dockerized services, health checks, and automated testing.

TECHNICAL SKILLS

AI/ML: LangChain, LangGraph, LlamaIndex, RAG, Prompt Engineering, LLMs, PyTorch, NLP

Languages: Python, Go, Java, C/C++, TypeScript, SQL

Frameworks & Data: Django, Flask, FastAPI, React, Kafka, Spark, Databricks, PostgreSQL, MongoDB, Redis

Cloud & DevOps: AWS, Azure, GCP, Docker, Kubernetes, CI/CD, Prometheus, Grafana

Certifications: NVIDIA AI Agents (Multimodal), DeepLearning.AI ML Specialization, Microsoft Azure Fundamentals