**KIRAN DAVULURI** California, USA | +1 (313) 641-2498 | dkiran238@gmail.com | https://www.linkedin.com/in/kiran-davuluri/ | https://github.com/kiran-collab

## CAREER SUMMARY

AI Engineer with 4+ years of experience designing and deploying machine learning and Generative AI solutions across financial services, healthcare, and enterprise automation. Experienced in building LLM applications, agentic AI systems, Retrieval-Augmented Generation (RAG) pipelines, NLP solutions, and scalable MLOps workflows on AWS and Google Cloud. Proficient in Python, PyTorch, LangChain, LangGraph, Vertex AI, and FastAPI, with expertise in model deployment, evaluation, governance, and production-grade AI systems that deliver secure, reliable, and business-focused outcomes.

## CORE SKILLS

**Programming Languages:** Python, SQL, Java, C++, C#, R, MATLAB

**Generative AI & Agentic AI:** LangChain, LangGraph, Retrieval-Augmented Generation (RAG), Prompt Engineering, OpenAI GPT-4, Claude, Gemini, Llama 3, Mistral, AutoGen, CrewAI, Model Context Protocol (MCP), Agent Development Kit (ADK), ReAct, Tool/Function Calling, Multi-Agent Orchestration, Langfuse, n8n

**Machine Learning & Deep Learning:** PyTorch, TensorFlow, Scikit-learn, XGBoost, LightGBM, Transformers, BERT, BioBERT, LSTM, MLP, Hyperparameter Tuning, A/B Testing

**Natural Language Processing (NLP):** Hugging Face Transformers, spaCy, NLTK, TF-IDF, Named Entity Recognition (NER), Text Classification, Sentiment Analysis

**Vector Databases & Search:** Pinecone, FAISS, Hybrid Search (Dense + BM25), Cross-Encoder Reranking

**LLM Fine-Tuning & Optimization:** PEFT, LoRA, QLoRA, Quantization, ONNX Runtime

**Cloud & MLOps:** AWS (Bedrock, SageMaker, Lambda, S3, EKS, EMR), Google Cloud Vertex AI, Azure AI, Docker, Kubernetes, MLflow, Apache Airflow, Apache Kafka, PySpark, Databricks, Terraform, Jenkins, CI/CD, CloudWatch

**Data Engineering:** Apache Spark, PySpark, Databricks, Apache Kafka, Airflow, Feast

**Model Evaluation & Governance:** SHAP, LIME, Evidently AI, Responsible AI, Bias Monitoring, LLM Evaluation

**Backend & APIs:** FastAPI, REST APIs, Git, GitHub Copilot, Claude Code, Codex, Agile/Scrum

## PROFESSIONAL EXPERIENCE

### AI Engineer | Infiswift.ai – California, USA — Nov 2025 – Present

- Engineered a modular NLP extraction API that consolidated multiple machine learning repositories into a unified inference pipeline, eliminating duplicate LLM workflows and improving maintainability across engineering teams.
- Developed a content-aware document validation agent using Vertex AI Gemini, LangChain, and LangGraph to accurately classify uploaded RFI documents based on content, reducing downstream processing failures caused by incorrect file categorization.
- Implemented a centralized FileType glossary injection mechanism for LLM prompts, enabling instant business rule updates without model retraining while ensuring consistent document classification across environments.
- Built a self-learning classification framework that automatically generated and validated regex patterns from newly identified document titles, expanding deterministic classification coverage while minimizing manual rule creation.
- Designed a scalable AWS serverless orchestration pipeline using Lambda, SQS FIFO, and Amazon S3 to manage automated rule promotion, deduplication, version control, and deployment across Development, Preview, and Production environments.
- Optimized wake-word detection models using multi-layer perceptron (MLP) architectures on Mel-spectrogram features, exporting optimized models with ONNX and TensorRT to enable low-latency edge inference for robotic arm control.
- Developed comprehensive unit testing and evaluation frameworks using PyTest and Moto to validate AI workflows, cloud integrations, and model behavior, improving deployment reliability and reducing regression risks.
- Conducted iterative LLM evaluation and A/B testing across multiple production projects, analyzing model agreement and reasoning quality to identify classification gaps, refine prompt strategies, and improve overall prediction consistency.

### AI/ML Engineer | JPMorgan Chase & co. – California, USA — Sept 2023 - Oct 2025

- Developed a real-time fraud detection platform processing 5M+ daily card transactions using LightGBM and Isolation Forest models, reducing confirmed fraud losses by 22% while improving transaction risk scoring.
- Built a Feast-based feature store integrated with Apache Kafka to serve streaming features with sub-50 ms latency, enabling real-time fraud detection and decisioning.
- Designed and implemented a GenAI-powered fraud investigation assistant using LangChain, Pinecone, and Retrieval-Augmented Generation (RAG), reducing analyst investigation time by 40% through contextual case retrieval.
- Trained transformer-based anomaly detection models in PyTorch to identify abnormal transaction patterns, increasing fraud detection accuracy by 15% while maintaining false positives below 2%.

Automated model training, validation, versioning, and deployment using Kubeflow Pipelines, Terraform, and AWS EKS, improving deployment consistency and accelerating model release cycles.

Implemented SHAP and LIME model explainability to provide transparent fraud risk predictions and support regulatory compliance for model governance and audit requirements.

Optimized production inference using ONNX Runtime and post-training quantization, improving scoring throughput by 3× while reducing inference latency for real-time transaction processing.

Collaborated with Risk, Compliance, and Data Science teams to implement responsible AI controls, including bias monitoring, output validation, and prompt injection safeguards, enabling secure production deployment of GenAI capabilities.

**Machine Learning Engineer | Cognizant – India** **Jan 2019 - Oct 2020**

Developed an XGBoost-based patient readmission prediction model using structured EHR data from 2M+ patient records, achieving 0.87 AUC-ROC and enabling care teams to identify high-risk patients before discharge.

Engineered scalable PySpark data pipelines on Databricks to process laboratory results, vital signs, and ICD-10 diagnosis data, reducing feature generation time by 60% for daily model refreshes.

Built an NLP pipeline using spaCy and BioBERT to extract symptoms, medications, and comorbidities from unstructured physician notes, improving structured clinical data coverage by 35%.

Established model drift and data quality monitoring with Evidently AI, enabling early detection of data distribution changes and supporting automated, HIPAA-compliant model retraining.

Deployed machine learning models as REST APIs using FastAPI, Docker, and AWS SageMaker, integrating with hospital EHR platforms through HL7/FHIR interfaces for real-time risk scoring.

Designed clinician-facing SHAP-based explainability dashboards to highlight key patient risk factors, improving model transparency and increasing physician adoption across pilot deployments.

Streamlined model lifecycle management using MLflow and Jenkins, automating experiment tracking, model versioning, and deployment while reducing release cycles from two weeks to three days.

Partnered with clinical and analytics teams to evaluate model-guided care interventions through an A/B test, contributing to an 18% reduction in 30-day patient readmissions during the pilot program.

## PROJECTS

**Built a multi-agent invoice extraction system** using Google Agent Development Kit (ADK) with an Orchestrator, File, and Extraction agent architecture to automate invoice processing. Leveraged Model Context Protocol (MCP) for standardized file access and implemented Pydantic-based validation with a trace-driven evaluation framework to measure extraction accuracy, hallucination rates, and workflow reliability while supporting secure, quality-gated deployments.

**Built a multi-agent compliance auditing solution** using the Claude Agent SDK, orchestrating five specialized agents with an MCP-based tool server to automate risk assessment and citation validation. Implemented guardrails, LLM evaluation pipelines, and production-grade CI/CD with security scanning and test coverage, delivering accurate, traceable, and audit-ready compliance reviews.

## EDUCATION

**Master of Science in Artificial Intelligence -** University of Michigan, Ann Arbor, Michigan, USA

**Bachelor of Technology in Information Technology (Computer Science) -** NMIMS University, Mumbai, Maharashtra, India

## PUBLICATIONS

**Griffin** - Facilitating real-time road event detection using UAV and GV (under review)

**BMEye** - Public Health-Oriented Body Mass Index Monitoring using Commodity WiFi (IEEE WFIoT '23)

**SWiDir** - Enhancing Smartphone-based Walking Direction Estimation with Passive WiFi Sensing (IEEE MASS '23)

**Minerva CQ Agent Assist** - peer-reviewed case study on a real-time Agent Assist product in voice-based customer support (co-author)