HARI SANKARA KOTESWARA LAL KELUTH

+1 (940)-843-7985 | harisankara.keluth@gmail.com | LinkedIn | Github | Portfolio ◊ **WILLING TO RELOCATE**

EDUCATION

University of North Texas,
MS in Computer and Information Science, GPA: 3.6/4.0 — Aug 2023 – May 2025

VR Siddhartha Engineering College
BTech in Computer Science Engineering, GPA: 3.5/4.0 — July 2019 – July 2023

PROFESSIONAL EXPERIENCE

AI/ML Intern | GrantAIde | Hybrid (USA/ Remote) — Apr 2025 – Present

- Designed and deployed Retrieval-Augmented Generation (RAG) pipelines using LangChain, FAISS, OpenAI GPT-4, and Google Gemini, enabling intelligent grant writing, document-aware semantic search, and LLM-based response generation.
- Built production-grade Flask APIs integrated with Firebase Firestore, Stripe Connect, and Google Cloud Run, delivering real-time AI services for summarization, entity extraction, chat history, and PDF reasoning workflows.
- Engineered scalable LLM pipelines with PyTorch, Hugging Face Transformers, and Google Generative AI Embeddings to automate document classification and improve semantic match accuracy by 40%, supporting a multi-tenant AI Copilot system.

Machine Learning Engineer | Cadential Technologies Pvt Ltd | India — *Jan 2022 – Jul 2023*

- Built a scalable fuzzy matching system using PySpark and Elasticsearch with Jaro-Winkler logic to process 6.4M records, enhancing data quality and consistency through a master data management framework.
- Designed and deployed machine learning models for real-time predictions using Python, TensorFlow, PyTorch, and Scikit-learn, supporting dynamic inference pipelines and boosting operational decision-making efficiency.
- Engineered NLP-driven search and recommendation systems using BERT, GPT, and Hugging Face Transformers, increasing semantic match accuracy by 35% and enhancing user experience across large-scale queries.
- Developed deep learning pipelines for image and text classification using CNNs, RNNs, and attention mechanisms, enabling scalable training and deployment for high-accuracy computer vision and language tasks.
- Implemented full MLOps deployment of ML models using Docker, Kubernetes, and CI/CD pipelines on AWS; conducted A/B testing and tuning to enhance model robustness and production performance.

AWS Cloud Intern | AICTE | India — *Oct 2021 – Dec 2021*

- Optimized scalable ETL workflows using AWS Glue, S3, Redshift, and Lambda, significantly improving data ingestion speed, transformation accuracy, and reducing pipeline latency across large-scale distributed environments.
- Built and deployed machine learning models using AWS SageMaker, AutoML, and EC2; designed end-to-end cloud pipelines with AWS Data Pipeline, Athena, and DynamoDB to enable efficient training and real-time data access.

Projects

ChatPDF using Retrieval Augmented Generation (RAG) *(GPT-3.5, FAISS, Hugging Face, OpenAI, NLP)*

- Developed an AI-powered PDF chatbot using RAG architecture for contextual search and semantic content extraction, enabling accurate question-answering on unstructured documents.
- Leveraged FAISS for vector indexing and GPT-3.5 Turbo for response generation, enhancing retrieval precision and language understanding in document-centric queries.

Insurance Cross-Sell Prediction *(FastAPI, Docker, AWS ECS, MLflow, CI/CD, DVC, Evidently AI)*

- Built a production-ready MLOps pipeline for cross-sell prediction including data ingestion, transformation, DVC-based versioning, and MLflow-driven experiment tracking to ensure performance reproducibility.
- Deployed the model with FastAPI on Docker inside AWS ECS, integrated CI/CD via GitHub Actions, and used Evidently AI for real-time drift detection and monitoring in production.

SKILLS

- **Machine Learning & AI:** Deep Learning, NLP, Computer Vision, LLMs (GPT-4, Gemini), RAG, LangChain, FAISS, TensorFlow, PyTorch, Scikit-learn
- **Data Engineering & Big Data:** SQL, MySQL, MongoDB, Snowflake, ETL, Data Pipelines, Hadoop, PySpark
- **Cloud & MLOps:** AWS (Glue, Redshift, EC2, SageMaker, Lambda), GCP (BigQuery, Dataflow, Vertex AI), CI/CD, Kubernetes, Docker, Terraform
- **Software Development:** Python, Java, C, React.js, Flask, Node.js, REST APIs, GraphQL, Model Deployment
- **Data Analytics & Optimization:** Tableau, Power BI, Feature Engineering, A/B Testing.

Certifications

Google Professional ML Engineer | **AWS Solutions Architect – Associate** | **AWS Cloud Foundations**