HARI SANKARA KOTESWARA LAL KELUTH
+1 (940)-843-7985 | harisankara.keluth@gmail.com | LinkedIn | Github | Portfolio | **WILLING TO RELOCATE**

SUMMARY

AI/ML Engineer with hands-on experience in **LLM fine-tuning, RAG-based chatbots, and MLOps pipelines**. Skilled in **PyTorch, Hugging Face Transformers, AWS, and scalable NLP solutions**. Proven ability to build production-grade AI systems with efficient inference, model optimization (LoRA, quantization), and automated CI/CD workflows.

EDUCATION

University of North Texas,
MS in Computer and Information Science, GPA: 3.5/4.0 — Aug 2023 – May 2025

VR Siddhartha Engineering College
BTech in Computer Science Engineering, GPA: 3.4/4.0 — July 2019 – July 2023

PROFESSIONAL EXPERIENCE

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 Scikitlearn, 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

Fine-Tuning LLaMA 2 on Custom Dataset *(LLaMA 2, LoRA, BitsAndBytes, Hugging Face, NLP, PyTorch, Transformers)*

• Fine-tuned LLaMA 2 (7B) using LoRA with 4-bit quantization to optimize memory and performance, enabling accurate Q&A responses on domain-specific wildfire datasets.

• Developed an inference pipeline by merging LoRA weights with LLaMA 2, leveraging Hugging Face Trainer and tokenizer to enhance contextual language understanding.

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.

Technical Skills

- **Machine Learning & AI**: Deep Learning, NLP, Computer Vision, LLMs (GPT-4, Gemini, LLaMA, Hugging Face Transformers), RAG, LangChain, FAISS, PyTorch, TensorFlow, Scikit-learn, LLM Fine-Tuning (LoRA, PEFT, Quantization).
- **Data Engineering & Big Data:** SQL, MongoDB, Snowflake, ETL, Data Pipelines, PySpark
- **Cloud & MLOps**: AWS (Glue, Redshift, SageMaker, Lambda, EC2), GCP, CI/CD, Kubernetes, Docker, MLflow, DVC
- **Software Development**: Python, Java, C, Flask, FastAPI, React.js, Node.js, REST APIs
- **Data Analytics & Optimization:** Tableau, Power BI, Feature Engineering, A/B Testing, Model Evaluation & Tuning

Certifications

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