Hari Keluth Ai/Ml Engineer

harikeluth7@gmail.com | (940) 843-7985 | USA | LinkedIn | GitHub | Portfolio

Summary

AI/ML Engineer with 4+ years of strong expertise in designing, developing, and deploying machine learning, deep learning, and generative AI solutions. Skilled in NLP, computer vision, ensemble models, and hybrid AI systems. Experienced with cloud platforms, MLOps pipelines, and scalable APIs. Adept at feature engineering, model evaluation, and delivering efficient, explainable, and production-ready AI solutions.

Technical Skills

- **Programming & Scripting**: Python, SQL, R, Java, C++, Bash
- **Machine Learning & Deep Learning**: TensorFlow, PyTorch, scikit-learn, XGBoost, LightGBM, CatBoost, GANs, K-Nearest Neighbors, Ensemble Learning, Transfer Learning, Reinforcement Learning, NLP, Computer Vision, Hybrid Tabular-Text Models, Supervised/Unsupervised Learning, Feature Engineering, Model Evaluation & Monitoring, Model Bias & Ethics
- **Generative AI & LLMs**: Hugging Face Transformers, GPT, BERT, LLaMA, Prompt Engineering & Chaining, RAG, LangChain, LlamaIndex, Vector Databases (FAISS, Pinecone, Weaviate), Agentic AI Frameworks, Fine-tuning & Alignment
- **Data Engineering & Big Data**: Pandas, NumPy, Apache Spark, Apache Kafka, Azure Databricks, Hadoop, MongoDB, Cassandra, Data Preprocessing & Cleaning, SQL/NoSQL Query Optimization, Working with Unstructured Data
- **Cloud Platforms & AI Services**: AWS SageMaker, AWS Lambda, AWS S3, Azure ML, Google Cloud AI Platform, Vertex AI, Cloud-native AI Solutions
- **MLOps & Deployment**: Docker, Kubernetes, MLflow, CI/CD, Jenkins, Terraform, Airflow, SageMaker Pipelines, FastAPI, REST & GraphQL APIs, Model Drift & Latency Monitoring (Prometheus, Grafana), End-to-End ML Pipeline Development, Scalable Microservices
- **Data Visualization & BI**: Tableau, Power BI, Matplotlib, Seaborn, Streamlit, Gradio
- **Model Evaluation & Optimization**: Cross-validation, Grid/Random Search, Bayesian Optimization, Hyperparameter Tuning, Precision, Recall, F1-score, ROC-AUC, Model Compression (ONNX, Quantization, Pruning, Distillation), Explainable AI

Professional Experience

AI/ML Engineer, *Fisher Investments* 10/2024 – Present | Remote, USA

- Worked on a Generative AI–driven investment recommendation platform, collaborating with data, product, and client advisory teams to deliver personalized portfolio suggestions and risk insights, resulting in 20% higher advisor adoption and 15% increase in client portfolio engagement.
- Integrated explainable AI modules using LLM embeddings, prompt engineering, and RAG, improving transparency of investment reasoning, which reduced client query resolution time by 25% and increased recommendation trust scores by 18% in advisor surveys.
- Engineered multimodal financial features from transaction history, risk profiles, market data, and textual research reports using transformer encoders and generative embeddings, improving recommendation accuracy by 14% and risk-adjusted portfolio performance (Sharpe ratio) by 12%, while mitigating allocation bias across client risk segments.
- Developed hybrid AI architectures combining generative models, ensemble learners, and transformer-based predictors, validated via cross-validation and fairness tests, achieving a 17% uplift in predicted portfolio returns and robust generalization across 95% of client segments.
- Designed and deployed scalable RESTful APIs using FastAPI, Docker, and SageMaker, integrated with MLflow-driven CI/CD pipelines, delivering low-latency (<200ms) recommendation services with 99.8% uptime, supporting real-time advisor dashboards and client-facing portfolio insights.

AI/ML Developer, *Novartis India* 07/2020 - 07/2023 | Vijayawada, India

- Designed and developed an advanced drug safety and adverse event detection system, collaborating with data scientists, analysts, and business stakeholders, aligning technical solutions with business goals, improving operational efficiency, and increasing early detection of adverse events by 25% across multiple drug portfolios.
- Used deep learning, transformer-based architectures, GANs, and XGBoost ensemble models using Azure ML Studio, performing hyperparameter tuning via grid search and Bayesian optimization, achieving 25% gain in detection of adverse drug reactions and reducing false alerts by 15%.
- Engineered robust features from structured and unstructured data using Python, SQL, and Databricks, uniting NLP embeddings and semantic features from clinical notes and patient reports, improving model detection accuracy by 20% and enhancing identification of subtle safety signals.
- Validated models using cross-validation, stratified sampling, and metrics including precision, recall, F1-score, and ROC-AUC, ensuring unbiased, production-ready performance, keeping 90% model accuracy and reducing operational risks from drug safety incidents.
- Deployed drug safety detection models on AWS SageMaker with automated pipelines via Lambda and S3, enabling real-time monitoring, continuous feedback loops, and improving detection speed by 30%, while keeping steady accuracy across multiple drug categories.
- Applied transformer-based NLP and sequence models to analyze clinical trial reports, patient feedback, and regulatory submissions, extracting semantic insights and hidden patterns, improving detection of adverse events, enhancing pharmacovigilance, and supporting faster decision-making.

Education

University of North Texas, Denton, TX, USA
Master of Science in Computer and Information Science 08/2023 – 05/2025
VR Siddhartha Engineering College, Vijayawada, India
Bachelor of Technology in Computer Science Engineering 07/2019 – 05/2023

Certificates

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