PURNA CHANDER KONDA

Saint Louis, MO | +1 (618)-974-7219 | kondapurnachander42@gmail.com | LinkedIn | Github | Portfolio

PROFESSIONAL SUMMARY

Results-driven AI/ML Engineer with 4+ years of experience designing and deploying production-ready ML models and scalable AI systems across finance, retail, and healthcare. Expert in Generative AI, RAG, and MLOps using Python, PyTorch, TensorFlow, Hugging Face, LangChain, and AWS. Skilled in NLP, Computer Vision, and Predictive Analytics, with a proven track record of automating workflows, improving efficiency, and delivering secure, high-impact business outcomes. Committed to Responsible AI and regulatory compliance.

TECHNICAL SKILLS

- **Programming & Scripting**: Python, SQL, JavaScript, R, Java
- **Machine Learning & Evaluation**: Supervised/Unsupervised Learning, Random Forest, XGBoost, LightGBM, CatBoost, SVM, Time-Series, Prophet, LSTM, Graph ML, GNN, Feature Engineering, Pandas, NumPy, Scikit-learn, Metrics, F1, AUC-ROC, BLEU, ROUGE
- **Large Language Models & Deep Learning**: PyTorch, TensorFlow, Keras, Hugging Face Transformers, GPT, BERT, T5, CLIP, Diffusion Models, GANs, VAEs, Fine-Tuning, LoRA, QLoRA, PEFT, NLP, spaCy, NLTK
- **Generative AI & Prompt Engineering**: OpenAI API, LangChain, LlamaIndex, Cohere, Anthropic Claude, DreamBooth-, Zero/Few-Shot Prompting, Chain-of-Thought, Retrieval-Augmented Generation, RAG, RAGAS, LangSmith
- **MLOps & Model Lifecycle**: MLflow, Kubeflow, KServe, AWS SageMaker, Vertex AI, DVC, Weights & Biases, CI/CD, Jenkins, GitHub Actions, Large-Scale ML Systems, A/B Testing, Optuna, Hyperparameter Tuning
- **Data Engineering & Pipelines**: Kafka, Apache Spark, Apache Airflow, BigQuery, Snowflake, Apache NiFi, ETL Pipelines, Data Warehouses, Distributed Computing
- **Vector Databases & Storage**: FAISS, Pinecone, Weaviate, ChromaDB, Milvus, PostgreSQL, MongoDB, Neo4j
- **Deployment & Cloud**: AWS, Bedrock, S3, Lambda, GCP, Vertex AI, Azure, Docker, Kubernetes, ONNX, TensorRT, Model Quantization, GGUF, AWQ, GPTQ, FastAPI, Flask, gRPC
- **Security, Responsible AI & Compliance**: SHAP, LIME, Fairlearn, Federated Learning, Responsible AI Standards, Model Governance, GDPR, HIPAA, FINRA, FCRA, CCPA, PII Protection
- **Monitoring & Visualization**: Prometheus, Grafana, ELK Stack, Model Drift Detection, Streamlit, Gradio, Tableau, Power BI

PROFESSIONAL EXPERIENCE

Broadridge **Apr 2025 - Present**

AI/ML Engineer *Kansas City, MO*

- Architected scaled AI/ML pipelines using RAG (LangChain, LlamaIndex) and vector stores (Pinecone, Weaviate, Milvus), boosting document analysis efficiency by 40% with RAGAS benchmarking for accuracy.
- Fine-tuned foundation models (GPT, LLaMA) using LoRA, QLoRA, PEFT; implemented Chain-of-Thought and RLHF to reduce hallucinations and improve compliance metrics by 30%.
- Orchestrated production-grade MLOps via Kubeflow, MLflow, and SageMaker Pipelines, maintaining 99.9% uptime and accelerating deployment cycles for financial services by 40%.
- Engineered high-throughput fraud detection and credit scoring pipelines using Apache Spark, Kafka, and Vertex AI, boosting anomaly detection by 25% through high-precision predictive modeling.
- Developed XAI dashboards with SHAP, LIME, and Fairlearn, enabling auditors to assess bias and meet strict FINRA, GDPR, and FCRA regulatory and compliance standards.
- Optimized model inference via ONNX and TensorRT with Quantization (GGUF/AWQ), reducing latency by 35% for high-traffic financial decision systems and real-time workflows.
- Constructed financial Knowledge Graphs with Neo4j, mapping complex relationships to uncover hidden fraud rings and enabling analysts to retrieve risk-linked insights 50% faster.
- Integrated AI microservices into enterprise platforms using AWS Bedrock, FastAPI, and gRPC, enabling secure, real-time financial workflows for 15K+ active enterprise users.

O'Reilly Auto Parts **Mar 2024 - Mar 2025**

Machine Learning Engineer *Springfield, MO*

- Architected demand forecasting engines using XGBoost and Prophet via PyTorch, optimizing regional inventory with DVC versioning to reduce stockout events by 18% across 1,200 hubs.
- Engineered ETL/ML pipelines via Apache Spark, Airflow, and Snowflake, orchestrating 100TB+ of transactional data for production-grade supply chain predictive analytics.
- Orchestrated containerized inference microservices using Docker and Kubernetes, maintaining sub-200ms latency for pricing models via Terraform-managed elastic infrastructure and Helm deployments.
- Streamlined MLOps frameworks via AWS SageMaker, Optuna, and MLflow, automating hyperparameter optimization and model tracking to accelerate retraining cycles by 35%.

- Constructed real-time streaming architectures using Kafka and Apache Flink for proactive anomaly detection, identifying logistics disruptions across 28 distribution centers with PySpark integration.
- Fine-tuned BERT-based NLP models to categorize 500,000+ unstructured parts, utilizing Temporal Fusion Transformers and Hugging Face to capture seasonal patterns and reduce overstock by 10%.
- Automated enterprise CI/CD workflows via GitLab CI and GitHub Actions, facilitating seamless model regression testing to maintain 99.9% uptime for mission-critical retail analytics microservices.
- Developed dashboards using Prometheus, Grafana, and Tableau, translating complex model telemetry into actionable supply chain KPIs that accelerated operational decision cycles by 30%.

Sahrudaya Healthcare **Aug 2022 - Nov 2023**
Data Scientist *Hyderabad, India*

- Developed predictive risk models using XGBoost and LightGBM, increasing patient engagement by 22% and reducing hospital readmissions by 15% through advanced clinical feature engineering.
- Fine-tuned BERT and RoBERTa NLP models via Hugging Face to classify 1M+ medical records, improving classification accuracy by 28% and reducing manual clinician review time by 40%.
- Architected HIPAA-compliant Spark and Hive ETL workflows on Azure and GCP (Vertex AI), processing high-volume clinical data while ensuring secure PII handling and 99.9% data reliability.
- Engineered Computer Vision pipelines using CNNs and OpenCV for automated medical imaging analysis, improving diagnostic screening speed by 30% and assisting in early pathology detection.
- Constructed time-series forecasting models using LSTM and Keras to predict emergency room capacity, improving patient throughput by 20% via TensorFlow-based deep learning.
- Developed model explainability frameworks using SHAP and MLflow, implementing automated drift detection to validate fairness and optimize patient prioritization via A/B testing.

Citco **Mar 2021 - Jul 2022**
Data Scientist *Hyderabad, India*

- Architected GNN-based fraud detection models and link analysis pipelines using Neo4j, uncovering hidden fraud rings across policyholder data and resulting in a 27% reduction in false positives.
- Developed high-performance time-series ensembles using CatBoost, Prophet, and LSTM, improving underwriting accuracy for insurance premium collections through automated hyperparameter tuning.
- Engineered scalable enterprise pipelines via Databricks and Spark MLlib, cutting reporting turnaround by 40% and enabling complex daily risk simulations on distributed datasets.
- Optimized distributed query performance by restructuring unstructured data pipelines in Cassandra and MongoDB, reducing retrieval latency for high-frequency financial markers by 40%.
- Deployed production ML microservices using Docker and Kubernetes, integrating Tableau/Power BI visual analytics to align model KPIs with global risk and compliance strategies.

EDUCATION

Webster University
Master's, information technology management *MO, USA*

CVR College of Engineering
Bachelor's, computer science and information technology *Hyderabad, India*

PROJECT HIGHLIGHTS

RAG-Based Research Paper Assistant
- Built a RAG-based research paper Q&A system using LangChain, HF embeddings, and FAISS, improving recall by 35%
- Deployed a FastAPI + Docker RAG API with Streamlit UI, enabling document ingestion and 1.2s query latency.

Diffusion-Based Image Generation Pipeline
- Built a Stable Diffusion and DreamBooth text-to-image pipeline for automated marketing asset generation.
- Trained models on AWS EC2 with PyTorch Lightning (DDP), reducing training time by 40% and deployed a Streamlit tool.

Explainable AI (XAI) Credit Risk Framework
- Built a Transformer-based credit scoring model with SHAP explainability, improving audit efficiency by 25%
- Applied Fairlearn and LIME for bias mitigation and automated retraining using Kubeflow and Airflow.

CERTIFICATIONS

- AWS CERTIFIED DATA ENGINEER - ASSOCIATE
- MICROSOFT CERTIFIED: AZURE AI ENGINEER ASSOCIATE