Niraj Koli

Pune, India +91-7709634340 nirajdkoli1234@gmail.com linkedin.com/in/nirajkoli
github.com/NKsCODES

Professional Summary

Python & AI/ML Engineer with **1.5+ years of experience** building production-ready ML pipelines, deploying deep learning models, and developing scalable AI applications using Python, TensorFlow/PyTorch, and cloud platforms. Skilled in **data preprocessing, model training, REST API integration (FastAPI/Flask), and MLOps workflows (MLflow, Docker, CI/CD)**. Experienced with **LLMs, NLP, Computer Vision, and Generative AI systems**.

Core Skills

- **Programming & Tools:** Python 3.x, SQL, Git, Jupyter Notebooks
- **Machine Learning:** scikit-learn, XGBoost, NumPy, Pandas, Matplotlib, classification, regression, clustering, feature engineering
- **Deep Learning:** PyTorch, TensorFlow, CNNs, Transformers, LLMs
- **Backend & Deployment:** FastAPI, Flask, REST APIs, Docker
- **Cloud & MLOps:** AWS (SageMaker/EC2), Azure AI, GCP, MLflow, CI/CD, monitoring, model versioning, A/B testing
- **Nice to Have:** NLP, Computer Vision, Time-Series Forecasting, ETL Pipelines, Airflow, Spark (exposure)

Professional Experience

Machine Learning Engineer (GenAI & Python Developer) **Apr 2024 – Mar 2025**
Pran's Rental Solutions Pvt. Ltd *Remote*

- Developed and maintained **Python-based ML applications**, including data ingestion, preprocessing, feature engineering, model training, and deployment.
- Built scalable **ML/DL pipelines** using PyTorch, scikit-learn, Transformers, and TensorFlow for NLP and recommendation tasks.
- Integrated trained models into **FastAPI microservices**, containerized with Docker, and deployed on AWS/Kubernetes.
- Optimized model performance using **benchmarking, A/B testing, hyperparameter tuning**, and production monitoring.
- Collaborated with data engineers & product teams to deliver reliable ML workflows with clean, well-documented code.

Selected Projects

Voice AI Pipeline (NLP + LLMs + FastAPI) **Oct 2025 – Present**

- Building a multilingual NLP-driven voice assistant using **Transformer-based LLMs** for real-time conversational interaction.
- Integrated **Azure STT/TTS**, REST APIs, and scalable FastAPI backend for low-latency responses.
- Experimenting with **optimization, latency reduction**, and multimodal expansion (CV hooks).

Automated ML & GenAI Pipeline (MLOps) [GitHub] **Aug 2025 – Sep 2025**

- Designed a full pipeline from **data ingestion → preprocessing → training → model deployment**.
- Implemented **model versioning, drift detection, monitoring**, and auto-retraining using MLflow.
- Designed workflow to be **Airflow/Spark-compatible**, aligned with enterprise-level MLOps practices.

Advanced Finance RAG Assistant (NLP + Vector Search) [GitHub] **Jun 2025 – Jul 2025**

- Developed an NLP-based RAG assistant using **FAISS vector search** over 50K+ documents.
- Improved accuracy by ∼25% using optimized retrieval, semantic embeddings, caching, and modular Python design.
- Demonstrated strong understanding of **NLP pipelines, semantic search, and model serving**.

Education

- **M.Sc. in Artificial Intelligence & Machine Learning** – Liverpool John Moores University & IIIT-Bangalore
  Expected Dec 2026
- **B.E. in Mechanical Engineering** – Savitribai Phule Pune University *2019 – 2023*

Certifications

- **Databricks Generative AI Fundamentals**
- **Microsoft Azure Data Fundamentals (DP-900)**