NITISH KUMAR MANTHRI

AI/ML Engineer | Generative AI | MLOps

nitish.m@protectmymails.com • (901) 699-6440 • Open to Relocate • LinkedIn • GitHub • Portfolio

PROFESSIONAL SUMMARY

AI/ML Engineer with 4.5+ years delivering production ML systems and GenAI applications generating measurable business impact. Expert in **LLM fine-tuning, RAG architectures, prompt engineering, agentic AI, and end-to-end MLOps**. Built enterprise platforms serving 90K+ analyst reviews with <2s latency and 99%+ uptime. Drove $250K+ cost savings and 15-35% efficiency gains across fraud detection, clinical NLP, and credit risk systems.

TECHNICAL SKILLS

ML & DL	Python, PyTorch, TensorFlow, Keras, XGBoost, LightGBM, CatBoost, Scikit-learn, CNNs, LSTMs, Transformers
GenAI & LLMs	GPT-4, LLaMA, Claude, Mistral, Fine-tuning (LoRA/QLoRA), Prompt Engineering, RAG, Agentic AI, LangChain, LlamaIndex
NLP & Search	BERT, Sentence Transformers, Hugging Face, NER, Pinecone, FAISS, ChromaDB, Weaviate, Semantic Search
MLOps & Cloud	MLflow, Docker, Kubernetes, FastAPI, CI/CD, AWS (SageMaker, Bedrock), Azure ML, Databricks, Airflow
Data & Eval	SQL, PySpark, Pandas, Kafka, A/B Testing, RAGAS, SHAP, Drift Detection, Hallucination Detection, Responsible AI

PROFESSIONAL EXPERIENCE

American Express | **AI/ML Engineer** | Aug 2024 – Present | USA

- Architected RAG-based AI Risk Copilot (LangChain, GPT-4, LLaMA) integrating transaction history and merchant risk signals; enabled 90K+ fraud/credit analyst reviews with full audit traceability
- Deployed Pinecone vector database indexing 7M+ records with BERT embeddings; achieved sub-100ms retrieval latency and reduced LLM hallucinations by 35% through semantic grounding
- Engineered multi-agent LLM workflows with chain-of-thought prompting for case summarization, anomaly reasoning, and automated policy citation; cut analyst review time by 40%
- Designed LLM evaluation framework (RAGAS metrics, hallucination detection, grounding checks) with human-in-the-loop validation; achieved 95% response accuracy pre-deployment
- Implemented Responsible AI controls (prompt injection safeguards, SHAP overlays, audit logging) ensuring Model Risk Management and regulatory compliance
- Built containerized microservices (Docker/Kubernetes) with <2s inference latency, 99.5% availability, and real-time drift monitoring (PSI/KL divergence)

CVS Health | **Machine Learning Engineer** | Sep 2023 – Jul 2024 | USA

- Programmed transformer-based clinical NLP (BERT/BioBERT) extracting structured indicators from 500K+ physician notes; achieved 90%+ F1-score for condition tagging
- Established predictive risk models (XGBoost, LightGBM) on 5M+ claims/EHR records; reduced high-risk cohort misclassification by 19% and prioritized 50K+ care interventions
- Configured A/B testing framework for pharmacy engagement (2M+ interactions); drove 15% improvement in medication adherence saving estimated $2M annually
- Formulated Azure Data Factory/Airflow ETL pipelines with automated validation; reduced data refresh latency by 30% while maintaining HIPAA compliance
- Rolled out models via MLflow with containerized APIs on Azure ML; implemented automated drift detection and retraining triggers

KPIT Technologies | **Data Scientist** | Jun 2020 – Aug 2022 | India

- Developed credit risk models (XGBoost, Random Forest) on 200K+ records achieving 78% precision; reduced false positive alerts by 15% saving $250K in manual review costs
- Orchestrated LSTM sequence models (TensorFlow) analyzing 12–18-month repayment histories; improved early default detection by 8% identifying $1.2M in at-risk accounts
- Formed behavioral features (rolling repayment ratios, utilization trends, delinquency indicators); improved high-risk account recall by 11%
- Created PyTorch NLP pipelines for customer service classification (0.83 F1); automated routing of 30K+ monthly dispute cases
- Released batch scoring APIs (FastAPI/Docker) with Airflow-orchestrated retraining and MLflow tracking; reduced model refresh cycle from 2 weeks to 3 days

EDUCATION

M.S. Computer Science — University of Memphis | May 2024 | Coursework: Deep Learning, NLP, Neural Networks, Big Data Analytics

PROJECTS

Clinical Decision Support Assistant: Multimodal clinical AI (GPT-4, LLaMA, BLIP-2) with RAG pipelines for medical image/text processing; 45% faster diagnostic review

Stock Price Prediction: Time-series forecasting with attention mechanism for S&P 500; integrated RSI, MACD, Bollinger Bands achieving 2.3% RMSE

Airline Customer Churn Prediction: LightGBM ensemble with SHAP explainability on 50K+ customers; 0.81 F1-score with SMOTE for class imbalance

Real-Time Object Detection for Autonomous Vehicles: YOLOv5 on KITTI dataset optimized with TensorRT; 78 mAP at 45 FPS for edge deployment

CERTIFICATIONS

AWS Certified ML Specialty • Azure AI Engineer Associate • Azure Data Scientist Associate • Google Cloud Professional ML Engineer • Databricks ML Associate • Databricks Data Engineer Associate • TensorFlow Developer Certificate • NVIDIA Certified Associate (GenAI & LLMs) • NVIDIA Certified Professional (GenAI)