RITHVIK REDDY PINNINTI

r.rithvik312@gmail.com | (806) 702-3056

6+ Years | Senior AI/ML Engineer | LLM / GenAI / MLOps / Cloud-Native ML Platforms

CAREER OBJECTIVE

Senior Machine Learning Engineer with **7+** years of experience building production **AI systems** and scalable ML infrastructure. Specialized in **LLM deployment, GPU-optimized inference, and cloud-native ML platforms** using **Python, Kubernetes,** and distributed computing. Proven track record of operationalizing models that serve high-throughput workloads with strong reliability and performance. Strong expertise in **LLM infrastructure, model hosting, fine-tuning (LoRA/QLoRA), RAG pipelines, embeddings,** and vector database integrations for real-time production systems. Experienced in taking **AI/ML and GenAI** solutions from proof-of-concept to production with focus on scalability, reliability, latency optimization, and cost efficiency. Skilled in building cloud-native AI platforms using **AWS, GCP, and Azure** with **Docker, Kubernetes, CI/CD pipelines,** and automated model monitoring. Hands-on experience with **MLOps**, experiment tracking, model registry, drift detection, and continuous retraining pipelines ensuring stable and reliable model performance.

PROFESSIONAL SUMMARY

- Senior **Machine Learning Engineer** with 6+ years of experience building production AI systems and scalable ML infrastructure across banking, retail, telecom, and insurance industries.
- Specialized in **LLM deployment, GPU-optimized inference,** and cloud-native ML platforms using Python, Kubernetes, and distributed computing with sub-second response times for high-volume enterprise workloads.
- Strong expertise in **LLM infrastructure, model hosting, fine-tuning (LoRA/QLoRA), RAG pipelines, embeddings,** and vector database integrations (Pinecone, FAISS, Chroma, Weaviate) for real-time production systems.
- Experienced in **Agentic AI workflows** using LangChain, AutoGen, CrewAI with ReAct patterns, multi-agent orchestration, tool calling, and memory-augmented agents for intelligent automation.
- Skilled in building cloud-native AI platforms using **AWS (SageMaker, Lambda, EC2, S3, ECS), GCP (Vertex AI, Dataflow, BigQuery), Azure ML** with Docker, Kubernetes, CI/CD pipelines, and automated model monitoring.
- Hands-on experience with **MLOps and LLMOps** including MLflow, Weights & Biases, experiment tracking, model registry, A/B testing, canary deployments, drift detection, and continuous retraining pipelines.
- Deep expertise in **Deep Learning frameworks** including PyTorch, TensorFlow, Keras, ONNX, TensorRT, Detectron2, YOLO, Transformers, mixed precision training, and distributed training.
- Strong **data engineering** skills with PySpark, Apache Spark, Apache Beam, Apache Airflow, Kafka, BigQuery, Snowflake, SQL, feature stores, and streaming/batch data processing.
- Proven ability to collaborate with **cross-functional product teams and ML research scientists** translating POC experiments into scalable production systems with load testing, security hardening, and compliance pipelines.

AI & GENAI CAPABILITIES

TECHNOLOGY	DEPLOYMENT	BUSINESS VALUE	TECHNICAL DEPTH
LLM Infrastructure & Hosting	Production GPU Clusters	Sub-Second Inference	Hugging Face, Triton Server, PyTorch, TensorRT, LoRA/QLoRA Fine-Tuning, Model Quantization, Tokenization Optimization
RAG & Vector Databases	Enterprise Production	Reduced Hallucinations	Pinecone, FAISS, Chroma, Weaviate, Embedding Pipelines, Sentence Transformers, Contextual Retrieval
Agentic AI & Orchestration	Workflow Automation	Intelligent Automation	LangChain Agents, AutoGen, CrewAI, ReAct Patterns, Multi-Agent Orchestration, Tool Calling, Memory-Augmented Agents
Predictive ML Models	3M+ Daily Predictions	Business-Critical Decisions	scikit-learn, XGBoost, LightGBM, Ensemble Methods, Classification, Regression, Time Series Forecasting, SHAP/LIME Explainability

NLP & Text Analytics	Multi-Domain	Actionable Insights	TF-IDF, Topic Modeling, Sentiment Analysis (VADER), NER, Text Classification, Gensim LDA, SentenceTransformers
Computer Vision	Real-Time Analytics	GPU-Accelerated	PyTorch, YOLO, Detectron2, TensorRT, ONNX, Mixed Precision Training, Distributed Training

INTEGRATION & MLOPS

TECHNOLOGY	LEVEL	SCOPE	TECHNICAL SKILLS
MLflow & Model Registry	Expert	End-to-End MLOps	Experiment Tracking, Model Versioning, A/B Testing, Canary Deployments, Blue-Green Releases, Automated Rollback
CI/CD for ML	Expert	Pipeline Automation	GitHub Actions, Jenkins, Docker Build, ECR Push, Linting, Unit/Integration Tests, Quality Gates
Kubernetes & Containers	Expert	Orchestration	Docker, K8s, StatefulSets, Autoscaling, GPU Scheduling, Terraform, Microservices Architecture
Monitoring & Observability	Expert	Production Reliability	Prometheus, Grafana, CloudWatch, Stackdriver, Drift Detection, Evidently AI, Alerting Systems
Data Pipelines	Advanced	Feature Engineering	PySpark, Apache Spark, Airflow, Beam, Kafka, BigQuery, Snowflake, Feature Stores, ETL/ELT

CORE SKILLS

DOMAIN	TOOLS & TECHNOLOGIES
LLM & Generative/Agentic AI	LLM Training, Fine-Tuning (LoRA, QLoRA, PEFT), RAG, Embeddings (Sentence Transformers, OpenAI), Tokenization, GPU Utilization, Inference Optimization, Triton Inference Server, Hugging Face Transformers, Vector Databases (Pinecone, Chroma, FAISS, Weaviate), Prompt Engineering, Model Quantization, LangChain Agents, ReAct, Multi-Agent Orchestration, Tool Calling, Memory-Augmented Agents
AI/ML Engineering	Regression & Classification Models, NLP (TF-IDF, Topic Modeling, Sentiment Analysis, NER), scikit-learn, VADER, Pandas, NumPy, PySpark ML, Feature Engineering, Time Series Forecasting, Model Explainability (SHAP, LIME), Experiment Tracking, Hyperparameter Tuning, Ensemble Methods, LightGBM, XGBoost
MLOps & LLMOps	CI/CD for ML, MLflow, Weights & Biases, Model Registry, A/B Testing, Canary Deployments, Blue-Green Deployments, Automated Retraining, Drift Detection, Performance Monitoring, GPU Scaling, Apache Airflow, Apache Beam, ETL Pipelines, Model Versioning, Shadow Testing, Model Risk Oversight
Cloud & DevOps	AWS (SageMaker, Lambda, EC2, S3, ECS), GCP (Vertex AI, Dataflow, BigQuery), Azure ML, Docker, Kubernetes, Terraform, FastAPI, Flask, Microservices Architecture, Event-Driven Pipelines, CloudWatch, Stackdriver, Prometheus, Grafana
Data Engineering	PySpark, Apache Spark, Apache Beam, Apache Airflow, Kafka, BigQuery, Snowflake, SQL, Data Warehousing, Streaming Data, Batch Processing, ETL/ELT, Data Quality, Feature Stores
Deep Learning	PyTorch, TensorFlow, Keras, ONNX, TensorRT, Detectron2, YOLO, Transformers, Mixed Precision Training, Distributed Training

PROFESSIONAL EXPERIENCE

Project 1: BMO Harris Bank (LLM Infrastructure & Enterprise AI Platform)

Role: AI/ML Engineer | Duration: ***Nov 2024 – Present***
Location: Chicago, IL
Environment: ***Hugging Face Transformers, Triton Inference Server, PyTorch, TensorRT, Docker, Kubernetes, Pinecone, Chroma, FAISS, MLflow, Prometheus, Grafana, FastAPI, AWS SageMaker, LangChain, CrewAI, AutoGen***

- Architected **GPU-accelerated LLM inference platform** using Hugging Face, Triton, and PyTorch, enabling sub-second response times for high-volume enterprise workloads across BMO's banking operations.
- Fine-tuned large language models **(LLaMA, Mistral)** using LoRA/QLoRA techniques, improving domain-specific accuracy while reducing training costs significantly for banking-specific use cases.
- Designed containerized deployment workflows with **Docker and Kubernetes**, implementing autoscaling and blue-green releases to ensure high availability and zero-downtime deployments.
- Built scalable **RAG pipelines** integrating vector databases (Pinecone/FAISS) to enhance contextual retrieval and reduce hallucinations in enterprise knowledge systems.
- Established **CI/CD pipelines for ML models** using GitHub Actions and MLflow, accelerating deployment cycles and improving model governance with automated versioning and rollback capabilities.
- Implemented real-time monitoring with **Prometheus and Grafana**, tracking latency, throughput, and GPU utilization to proactively detect performance bottlenecks and ensure SLA compliance.
- Developed secure **API-based model serving frameworks** with rate limiting, content filtering, and audit logging to meet enterprise compliance standards and regulatory requirements.
- Developed and optimized AI applications using **Python, LangChain, and CrewAI**, implementing Retrieval-Augmented Generation (RAG) and Agentic AI workflows for intelligent automation of banking processes.
- Developed autonomous agent workflows using **LangChain and AutoGen** to enable intelligent task automation with multi-agent orchestration, tool calling, and memory-augmented agents.
- Built scalable API services for serving **RAG-based generative AI solutions on AWS SageMaker** and deployed models for scalable inference and real-time predictions across multiple business units.
- Built centralized model management platform using **MLflow**, enabling automated versioning, A/B testing, and rollback for production reliability across all ML models in the organization.
- Developed detailed **technical documentation** including API specifications with OpenAPI schemas, architectural diagrams with Lucidchart, performance benchmarking reports, and operational runbooks covering incident response, scaling procedures, and troubleshooting workflows.
- Collaborated with **cross-functional product teams and ML research scientists** translating POC experiments into scalable production systems implementing load testing with Locust, security hardening with prompt injection detection, and PII redaction pipelines.

Project 2: BestBuy (ML Model Operationalization & Production Deployment)

Role: AI/ML Engineer | Duration: ***Aug 2023 – Nov 2024***
Location: Richfield, Minnesota
Environment: ***Python, PySpark, scikit-learn, VADER, Pandas, NumPy, TF-IDF, Gensim LDA, SentenceTransformers, MLflow, Apache Airflow, Docker, Kubernetes, FastAPI, Kafka, BigQuery, SHAP, Evidently AI, Databricks***

- Productionized **machine learning models** using scikit-learn and gradient boosting frameworks, supporting 3+ million daily predictions across business-critical retail applications including demand forecasting and customer analytics.
- Engineered distributed **ETL pipelines using PySpark and Databricks** to support large-scale feature engineering and automated retraining workflows processing terabytes of retail transaction data.
- Built RESTful inference services using **FastAPI**, enabling real-time and batch prediction capabilities for downstream systems with comprehensive error handling and request validation.
- Implemented automated **model monitoring with drift detection** and performance metrics using Evidently AI, improving model reliability and reducing production incidents across the ML fleet.
- Containerized ML workloads using **Docker** and deployed to Kubernetes, significantly improving scalability and deployment consistency across development, staging, and production environments.
- Developed **NLP pipelines** leveraging TF-IDF, embeddings, and NER to extract actionable insights from unstructured customer reviews and product descriptions at scale.
- Designed **A/B testing frameworks** to evaluate model performance with statistical significance testing, enabling data-driven release decisions for business-critical ML models.
- Built **automated monitoring pipelines** generating model performance metrics including AUC-ROC curves, precision-recall analysis, confusion matrices, statistical drift detection using KS tests and PSI calculations, feature importance tracking with SHAP values, and compliance dashboards for Model Risk Oversight teams.
- Automated end-to-end **CI/CD pipelines in Jenkins**: linting Python code, running unit and integration tests, building Docker images, and pushing to Amazon ECR with automated quality gates.

- Collaborated with **business stakeholders and data analysts** translating requirements into ML solutions implementing custom loss functions, business constraint optimization, and explainability reports using LIME local interpretations ensuring stakeholder confidence.

Project 3: Cisco (GPU-Accelerated ML Systems & Computer Vision)

Role: Python & ML Engineer | Duration: ***Aug 2020 – Jul 2022***
Location: San Jose, California
Environment: ***PyTorch, YOLO, TensorRT, Docker, Kubernetes, MLflow, PySpark, Apache Beam, Kafka, Redis, Prometheus, Grafana, FastAPI, RabbitMQ, Terraform, AWS EC2***

- Built **GPU-accelerated computer vision and ML inference systems** using PyTorch, YOLO, and TensorRT for real-time analytics across Cisco's enterprise networking and security product lines.
- Deployed containerized **ML pipelines on Kubernetes and AWS** improving scalability and reliability of AI workloads with auto-scaling, health checks, and resource optimization.
- Developed **machine learning models** for forecasting, classification, and recommendation systems across enterprise projects supporting network traffic analysis and anomaly detection.
- Built large-scale **Spark and Airflow data pipelines** for feature engineering and analytics across structured and unstructured datasets with automated scheduling and monitoring.
- Implemented end-to-end **MLOps** including model versioning, monitoring, validation, and automated retraining workflows using MLflow for experiment tracking and model registry management.
- Developed **NLP solutions** for sentiment analysis, topic modeling, and text classification improving business insights from customer feedback and support ticket analysis.
- Optimized inference performance using **quantization and batching techniques** reducing latency by 60% and compute costs by 40% for production workloads.
- Implemented monitoring **dashboards and alerting systems** using Prometheus and Grafana ensuring production reliability, uptime, and proactive incident detection.

Project 4: Intact Insurance (Scalable Backend & Data-Driven ML Services)

Role: Python Developer | Duration: ***Jan 2018 – Jul 2020***
Location: Chicago, IL
Environment: ***Python, SQL, PySpark, Pandas, NumPy, scikit-learn, TensorFlow, XGBoost, LightGBM, Docker, FastAPI, Flask***

- Developed scalable **backend applications and data-driven services** using Python, implementing modular architecture, reusable components, and robust error handling for production environments serving insurance claim processing.
- Designed and optimized high-volume **data pipelines using Python and Apache Spark**, processing large datasets while improving runtime performance through efficient memory management and parallel execution strategies.
- Built **RESTful APIs using FastAPI/Flask** to expose machine learning models and business logic, enabling real-time data access for downstream applications with comprehensive API documentation.
- Wrote complex **SQL queries** and integrated Python with relational databases to support high-performance data retrieval, transformation, and analytics for insurance underwriting models.
- Developed data processing utilities using **Pandas and NumPy** for cleansing, transformation, and feature engineering across multiple business datasets including claims, policies, and customer records.
- Containerized Python applications using **Docker** and supported CI/CD pipelines through Git-based version control and automated deployments with quality gates and testing stages.
- Integrated **cloud services** with Python applications for scalable compute, storage, and event-driven processing supporting insurance analytics workloads.
- Built **logging, monitoring, and alerting mechanisms** to proactively detect failures and maintain production reliability across all deployed services.
- Collaborated with **cross-functional teams** to translate business requirements into technical solutions, delivering production-grade software within agile environments with sprint planning and retrospectives.
- Improved application performance by profiling **bottlenecks, optimizing algorithms**, and implementing asynchronous processing where applicable, achieving 3x throughput improvements on key pipelines.

PRIORITY PROJECTS

LLM Infrastructure & Hosting Platform

Technologies: ***Hugging Face Transformers, Triton Inference Server, Kubernetes, Vector Databases, PyTorch, TensorRT***

- Hands-on **LLM training and fine-tuning implementations** using LoRA/QLoRA parameter-efficient techniques for domain-specific model adaptation
- Production **GPU-centric model server design** and Kubernetes orchestration with StatefulSets for persistent model state management
- Custom **embedding pipelines and tokenization optimization** using SentencePiece for domain adaptation with vocabulary extension
- Low-latency **API inference architecture with sub-second response times** using Triton Inference Server with dynamic batching and model ensemble
- Comprehensive **LLM observability dashboards** tracking performance metrics, token throughput, GPU utilization, and reliability KPIs

ML Model Operationalization -- Production Deployment

Technologies: ***Python, PySpark, scikit-learn, VADER, Pandas, MLflow, Docker, Kubernetes***

- Production **regression models and classification pipelines** with automated deployment, versioning, and rollback capabilities
- Advanced **NLP models** including TF-IDF vectorization, topic modeling with Gensim, and sentiment analysis pipelines
- Distributed **PySpark ML pipelines** for large-scale training on Databricks clusters processing terabytes of data
- Comprehensive **Model Risk Oversight dashboards** with drift detection using KS tests, PSI calculations, and feature importance monitoring
- Seamless **business integration of ML services** with REST APIs, batch processing, and event-driven architectures

CERTIFICATIONS

AWS Machine Learning -- Specialty -- Amazon Web Services
Google Cloud Machine Learning Engineer -- Google Cloud
TensorFlow Developer Certificate -- Google
PyTorch Developer Certification -- Meta

EDUCATION

Master’s / Bachelor’s in Computer Science / Engineering