Yajurved Jayavarapu

San Francisco, CA | +1 (205)259-6285 | yjayavarapu@gmail.com| LinkedIn | GitHub | Medium

Summary

AI Engineer specializing in Generative AI and production ML systems, with 4 years of experience designing and deploying scalable LLM platforms. Expertise in RAG architectures, multi-agent systems, and low-latency AI services on AWS Bedrock. Proven ability to build end-to-end AI systems that reduce operational cost, improve decision workflows, and scale reliably in production environments.

Skills

Programming: Python, SQL
Generative AI & LLMs: OpenAI, AWS Bedrock, RAG, LangChain, LangGraph, Prompt Engineering, Agentic Workflows
ML & AI: NLP, Deep Learning, Classical ML, Model Optimization (Quantization, Distillation)
LLM Systems: Embeddings, Vector Search, Retrieval Systems, Context Engineering, Multi-Agent Systems
Infrastructure: MLflow, Airflow, Docker, Kubernetes, CI/CD, Experiment Tracking
Cloud Platforms: AWS (Bedrock, SageMaker, Lambda, S3, SQS, Step Functions), Azure
Databases: PostgreSQL, MySQL, Vector DBs (FAISS, Pinecone)
APIs & Backend: FastAPI, REST APIs, Async Processing, Redis Caching
Monitoring & Governance: Model Monitoring, Drift Detection, AI Guardrails, NIST AI Risk Framework

Experience

Syneos Health

AI Engineer (Generative AI & ML PlatForms) ***Jan 2025 - Present***

- Leading development of enterprise-scale GenAI platform supporting claims processing, underwriting, and customer service automation
- Designed and deployed **RAG-based LLM systems** using AWS Bedrock, improving processing efficiency by 40%
- Architected modular LLM pipeline (embedding → retrieval → ranking → context assembly → inference) enabling scalable and reusable AI services
- Built **multi-agent workflows and internal AI assistants** using **LangGraph** and hybrid retrieval strategies
- Delivered **low-latency inference (<400ms**) through async batching, quantization, and Redis-based caching
- Developed intelligent document summarization and knowledge retrieval systems using FAISS and Pinecone
- Orchestrated end-to-end ML pipelines using **Airflow, MLflow, Docker, and Kubernetes** for production deployment
- Implemented real-time event-driven AI services using AWS Lambda, SQS, and Step Functions
- Improved retrieval relevance by **32% Recall@10**, reducing manual workload across support operations
- Built **GenAI guardrails** including hallucination detection, toxicity filtering, and audit logging for enterprise safety
- Ensured compliance with **NIST AI risk frameworks** and implemented monitoring and drift detection for production models

University of Alabama at Birmingham

Machine Learning Engineer (Data Systems & Analytics) ***Aug 2023 - Apr 2024***

- Built real-time data pipelines integrating multi-source sensor data, improving data accuracy by 25%
- Designed and deployed analytics systems reducing reporting latency by 60%
- Developed automated ETL pipelines and API integrations for continuous data ingestion and processing
- Delivered production-ready data systems enabling real-time decision-making for stakeholders
- Implemented data validation and anomaly detection checks, improving reliability of incoming data streams
- Optimized data processing workflows, reducing pipeline execution time and improving system throughput
- Collaborated with cross-functional teams to translate raw data into actionable ML-ready features

Thomson Reuters

Machine Learning Engineer(NLP & Data Platforms) ***Aug 2021 - Apr 2023***

- Developed and deployed scalable ML pipelines for NLP and analytics applications across large datasets
- Built distributed data processing systems using Apache Spark, enabling efficient model training at scale
- Automated ML workflows using Airflow, reducing pipeline latency by 35%
- Designed NLP systems for text classification, sentiment analysis, and information extraction
- Applied rigorous evaluation frameworks (AUC, F1, Precision/Recall) to ensure model performance in production
- Collaborated with engineering teams to productionize ML systems with reliability and scalability
- Engineered feature pipelines and reusable datasets consumed across multiple ML use cases
- Improved model performance through feature engineering and hyperparameter tuning, increasing predictive accuracy
- Implemented model versioning and experiment tracking using MLflow for reproducibility
- Built data preprocessing pipelines for large-scale unstructured text, improving downstream NLP model efficiency

Education

Master of Science, Data Science
University of Alabama at Birmingham | AL, USA | **May 2023 - Dec 2024**
Bachelor of Technology, Computer Science
SRM Institute of Science and Technology | INDIA **June 2018 - May 2022**