Yajurved Jayavarapu

San Francisco, CA | +1 (510)404-3675 | yjayavarapu@gmail.com| LinkedIn | Publication | GitHub | Medium

Summary

AI Engineer / Machine Learning Engineer with 3+ years of experience designing and shipping production LLM systems, RAG pipelines, and agentic AI workflows. Proven ability to architect end-to-end AI solutions—from data ingestion and feature engineering to LLM evaluation and system-level validation. Deep hands-on expertise in LangChain, LlamaIndex, LangGraph, vector databases, and scalable cloud-native ML platforms on AWS. Strong foundation in distributed computing (Spark), Python, and MLOps practices for reliable, production-grade AI systems.

Education

Master of Science, Data Science
University of Alabama at Birmingham | AL, USA | May 2023 - Dec 2024

Skills

- **AI / LLM Systems**: OpenAI API, Anthropic API, RAG, Agentic AI, LangChain, LangGraph, LlamaIndex, Prompt Engineering, Vector Databases (ChromaDB, FAISS), LLM Observability (LangSmith), Function Calling/Tool Use.
- **Machine Learning:** Classical ML, Time Series, Deep Learning (CNN, RNN), NLP, Feature Pipelines, Model Evaluation, Experiment Tracking
- **Data Engineering:** Apache Spark, Airflow, ETL Pipelines, AWS (S3, EC2, EMR, Redshift)
- **Programming & Tools:** Python, SQL, R | PostgreSQL, MySQL, SQL Server | Power BI, Tableau | Git, JIRA, Jupyter.

Experience

Syneos Health
Data Scientist | USA | Jan 2025 - Present

- Designed and owned end-to-end ML pipelines in Python and SQL, from ingestion and feature engineering to training and evaluation.
- Built Spark-based batch pipelines processing large-scale temporal datasets for predictive modeling and analytics.
- Orchestrated automated ETL and ML workflows using Airflow DAGs on AWS, reducing data processing latency by 35%.
- Developed reusable ML components and datasets consumed by multiple downstream analytics and AI workflows.
- Collaborated with data engineers and platform teams to ensure data quality, reproducibility and versioned ML workflows.
- Converted unstructured external research and reports into structured, machine-consumable datasets to support analytics and AI-driven decision systems.

University of Alabama at Birmingham
Data Scientist | AL, USA | Aug 2023 - Apr 2024

- Engineered real-time data ingestion pipelines from multiple sensor and API sources, improving data accuracy by 25%.
- Unified data from independent tracking systems into a single analytics layer for downstream modeling and visualization.
- Implemented data validation and anomaly detection logic, increasing reliability of performance metrics by 40%.
- Built interactive dashboards backed by standardized datasets, reducing manual reporting time by 60%.

Thomson Reuters
Data Scientist | INDIA | Aug 2021 - Apr 2023

- Built Spark-based ML pipelines on Hadoop clusters for large-scale data preprocessing, feature engineering, and modeling.
- Developed production-oriented ML workflows using Scikit-learn with reproducible training and evaluation steps.
- Implemented NLP pipelines for large-scale text processing and sentiment analysis.
- Applied cross-validation and robust metrics (AUC, F1, Precision, Recall) to validate ML systems.
- Partnered with data engineering teams to operationalize analytics and ML workflows.

Projects

Project1: Grounded RAG Assistant (LLMs, LangChain, ChromaDB)

- Designed a grounded RAG system enforcing document-only answers and safe refusals to prevent hallucinations.
- Built an agent-based query routing layer to constrain retrieval by domain while reusing a shared RAG pipeline.
- Implemented hybrid retrieval (BM25 + dense vectors) with metadata-aware filtering and chunk-level traceability.
- Developed debugging tools to inspect retrieved chunks, metadata filters, and failure cases for explainable RAG behavior.
- Added integration tests and optional tracing to validate end-to-end LLM and retrieval behavior.

Project2: Verified Agentic RAG System

- Designed a multi-agent retrieval-augmented generation (RAG) system with Planner, Retriever, Answer, and Verifier agents.
- Implemented query decomposition, multi-hop retrieval, citation enforcement, and retry logic.
- Enforced strict grounding—LLM responses generated only from retrieved documents.
- Built using modular agent architecture, state machines, and structured outputs (JSON schemas).