Akhilanand Benkal Venkanna

abenkalv@gmail.com | 312-536-0070 | Princeton, NJ | linkedin.com/in/abenkalv

Summary

Lead ML Engineer with 9+ years building production generative AI, machine learning, NLP, and data systems on AWS and Databricks. Specializes in AI agents, retrieval-augmented generation, text-to-SQL, and LLM evals.

Experience

Lead ML Engineer — 01/2022 – Present
Johnson & Johnson

- Shipped an agentic LangGraph assistant for 1,400 field reps and 50 regional managers, routing questions to text-to-SQL, hybrid RAG over Textract-extracted PDFs, SKU lookup, and account search; cut time-to-answer by ~30%.
- Orchestrated three Claude models on AWS Bedrock: Opus for accuracy-critical routing and multi-step tool calls, Sonnet for answers, and Haiku for query rewriting, serving ~2,000 queries a day through FastAPI.
- Improved routing accuracy from 80% to 95%+ through iterative error analysis; added LangGraph retry loops for failed SQL and clarifying questions when queries returned 100+ rows.
- Lifted RAGAS faithfulness from 0.75 to 0.9+ with top-k tuning, reranking, and source-only answering; used GPT judges for groundedness and guardrail adherence and traced live traffic in Databricks MLflow.
- Led a team of 6 delivering a pricing engine that recommends line-item discounts from price elasticity and optimization, realizing $4M a year by curbing excess discounts on new products; built its customer segmentation models.
- Designed a three-stage XGBoost and Random Forest pipeline classifying ~400K products from customer files and FDA GUDID, filtering out non-devices and mapping each to a J&J platform and segment at 95% and 92% accuracy.
- Cut analyst mapping time by 40% with a one-click shortlist of the 20 closest J&J SKUs, using TF-IDF to narrow candidates to 100 and a fine-tuned PubMedBERT model to rank them at 76% top-20 accuracy.

Data Engineering Senior Analyst — 10/2019 – 01/2022
Johnson & Johnson

- Built the daily Airflow and PySpark pipelines on AWS for a COVID-19 dashboard that forecast sales at the zip-code level from projected case spread and hospitalization trends.
- Migrated local data science workloads to EMR and SageMaker, rewriting complex joins in PySpark and tuning Spark jobs to cut runtimes by 60%.

Data Engineer — 02/2018 – 10/2019
Monsanto

- Engineered streaming ETL in Scala, Kafka, and Akka for the Customer360 asset platform, consuming 20+ source systems and processing millions of records.
- Developed RPC APIs in Go and deployed services as Docker containers on Monsanto's internal Kubernetes cluster.

Full Stack Senior Developer — 06/2017 – 02/2018
Colaberry

- Integrated Oracle, Teradata, SAP, and Salesforce into Monsanto's Customer360 platform through Kafka Connect and Scala Akka Streams ingestion services, containerized with Docker.

Skills

Languages: Python, SQL, Scala, Go

LLM: LangGraph, LangChain, prompt engineering, vector search, Claude, AWS Bedrock, OpenAI GPT, RAG, embeddings, rerankers, tool calling, Pydantic, guardrails, text-to-SQL

ML and Evaluation: LLM-as-judge, RAGAS, Databricks, MLflow, PyTorch, Sentence Transformers, embedding fine-tuning, XGBoost, NER, OCR

Data and Cloud: Spark, PySpark, Airflow, Kafka, AWS SageMaker, EMR, OpenSearch, Textract, FastAPI, Docker, Kubernetes, PostgreSQL

Education

Master of Science, Computer Science — 08/2015 – 05/2017
Illinois Institute of Technology — Chicago, IL