

Nikhil Reddy

(703)-906-6508 | nikhilreddyulindala@gmail.com | [LinkedIn](#) | [Portfolio](#) | [GitHub](#)

PROFILE SUMMARY

AI/ML Engineer with 3+ years building production LLM systems, RAG pipelines, and multi-agent applications across healthcare and enterprise. Track record of taking projects from prototype to production - fine-tuning BERT and Llama models, designing evaluation frameworks with RAGAS, and shipping systems that clinical and engineering teams use daily.

TECHNICAL SKILLS

Programming & Query Languages: Python, SQL, PySpark, R

ML & Deep Learning: PyTorch, TensorFlow, Scikit-learn, XGBoost, Hugging Face Transformers, BERT

GenAI & LLM Tooling: LangChain, LangGraph, RAG, Prompt Engineering, FAISS, Pinecone, ChromaDB, OpenAI API, Claude API,

Agentic AI: Multi-Agent Orchestration, Tool Calling, Human-in-the-Loop, CrewAI, AutoGen, MCP, LangSmith

Evaluation & Observability: RAGAS, DeepEval, LLM Evaluation, A/B Testing, Model Monitoring, Drift Detection

MLOps & Cloud: Docker, Kubernetes, MLflow, Airflow, FastAPI, CI/CD, AWS (SageMaker), Azure ML, GCP (Vertex AI)

Data & Databases: PostgreSQL, MongoDB, Snowflake, Databricks, Apache Spark, dbt, BigQuery

WORK EXPERIENCE

AI/ML Engineer – Milken Institute School of Public Health | Washington D.C

Jan 2025 – May 2026

- Built a LangChain RAG pipeline over 15K behavioral health records using GPT-4 and custom prompts; measured retrieval quality with RAGAS, achieving 34% improvement over baseline, enabling clinical teams to handle 3x more patient queries.
- Fine-tuned BERT via LoRA on 5K patient survey responses for clinical text classification, pushing accuracy from 71% to 87% on held-out test sets; worked directly with clinical staff to validate model outputs and document edge cases.
- Containerized FastAPI inference endpoints on AWS SageMaker with Docker; added input validation and output safety constraints for clinical decision-support workflows, cutting average response latency by 45%.

Data Scientist (NLP/ML) – DSSD | Washington D.C

Aug 2024 – Dec 2024

- Deployed FAISS semantic search with anomaly validators across JSON, CSV, and Parquet census data; partnered with the data engineering team to trace cross-format quality gaps and bring the pipeline to full compliance.
- Designed automated feature engineering workflows in PySpark on Databricks, using LangGraph agents to surface and test candidate features; increased statistical modeling throughput by 35% for a policy research team.
- Shipped Streamlit dashboards backed by Azure OpenAI; translated model outputs into plain-language summaries for non-technical leadership, reducing reporting prep time by 40%.

Data Scientist – Cogno AI | India

May 2022 – Jul 2024

- Built a production RAG system using OpenAI embeddings and Pinecone vector search across 10K e-commerce product SKUs; reduced average query latency from 3.2s to 0.8s, directly improving search relevance and customer conversion.
- Trained an XGBoost demand forecasting model on retail POS data, improving prediction accuracy 17% over an ARIMA baseline and reducing annual inventory overstock.
- Set up MLOps infrastructure using MLflow, Git, and Docker with experiment tracking, A/B model evaluation, and observability dashboards; cut the team's model iteration cycle by 45%.

PROJECTS

Autonomous Multi-Agent Research Assistant

- Built a multi-agent research assistant using LangGraph and CrewAI across 50K documents; agents plan queries, execute searches, and self-critique in loops - reaching 84% task completion accuracy on held-out evaluation queries.
- Tracked agent outputs with LangSmith and measured response quality with RAGAS; iterated on prompt design and retrieval routing, reducing hallucination rate by 38% and average response latency by 52%.

LLM Fine-Tuning & RAG Document Intelligence Platform

- Fine-tuned Llama-3 and Mistral via QLoRA on domain-specific corpora; connected Pinecone-backed retrieval across 23K enterprise documents, reaching 89% answer accuracy on a held-out evaluation set.
- Shipped RAGAS and DeepEval evaluation pipelines with A/B testing on AWS SageMaker; the framework surfaced only high-confidence answers to reviewers, reducing document review time by 70%.

Agentic Clinical AI System with Human-in-the-Loop Workflows

- Designed a clinical triage system on AWS Bedrock using LangChain tool-calling; added a human review gate before any automated action, safely handling 78% of patient record summarization tasks without requiring physician escalation.
- Improved clinical decision recall by 31% by connecting MCP with ChromaDB retrieval; monitored model behavior in production via MLflow, flagging drift and low-confidence outputs for human review.

EDUCATION

Master's in Data Analytics – The George Washington University

Bachelor of Technology in Computer Science and Engineering – Dayananda Sagar University