SAMHITA KOLLURI

(781) 947-7769 | samhita.kolluri@gmail.com | linkedin.com/in/samhita-kolluri | github.com/Samhita-kolluri | Hugging Face | samhita-kolluri.github.io

EXECUTIVE SUMMARY

Generative AI Engineer with 3+ years in Data Engineering, MLOps, and ML Systems. Expert in building Autonomous Agents and architecting RAG pipelines with LangChain and ChromaDB. Spearheaded multi-agent systems, fine-tuning, and inference optimization using Vector Databases, translating complex research into scalable, high-impact production-ready solutions.

SKILLS

Languages: Python, SQL, Bash
Generative AI & ML: LLMs (Fine-Tuning, RAG), LangChain, Agents, HuggingFace, Vector DBs, PyTorch, TensorFlow, OpenCV
Data Engineering & Ops: AWS, GCP, Databricks, Snowflake, Airflow, dbt, Docker, Kubernetes, CI/CD, Tableau, Power BI

WORK EXPERIENCE

Humanitarians AI | Boston, MA **January 2025 - May 2025**

AI Research and Engineer Co-op **Link**

- Architected a Contrastive Agent utilizing hybrid embeddings and stance-aware retrieval to autonomously identify semantically opposing documents within large unstructured corpora, establishing a foundational framework for automated narrative analysis.
- Fine-tuned a LLaMA-3 model on a custom contradiction dataset using parameter-efficient techniques (PEFT), significantly enhancing the model's ability to detect nuanced conflicts in text compared to off-the-shelf base models.
- Orchestrated a robust RAG pipeline with ChromaDB, implementing a two-phase retrieval strategy that re-ranks retrieved context based on relevance scores, directly improving the accuracy of conflicting narrative detection by 35%.
- Engineered persistent memory layers and observability tools using LangChain, enabling the tracking of agent reasoning traces across long-context interaction loops and ensuring reliable, interpretable decision-making in production.

Cognizant Technology Solutions | Hyderabad, INDIA **August 2022 - July 2023**

Senior AI Data Engineer

- Spearheaded the enterprise warehouse modernization for a Fortune 500 Insurance client, leading the strategic end-to-end migration of legacy Informatica workflows to optimized, scalable SQL layers to support high-volume real-time analytics.
- Re-architected critical reconciliation pipelines by implementing vectorized operations and advanced query optimization techniques, successfully reducing execution time from hours to minutes to achieve a 99% efficiency gain in processing.
- Designed and implemented automated audit frameworks and Python-based monitoring scripts, ensuring strict data governance compliance and proactively reducing production data quality incidents by 25% across the entire enterprise ecosystem.

AI Data Engineer **March 2021 - August 2022**

- Deployed to a specialized Generative AI Research Lab to engineer high-throughput PySpark pipelines on Databricks, sanitizing and tokenizing massive unstructured text corpora specifically for large-scale Large Language Model (LLM) pre-training.
- Designed complex distributed data transformation logic in Python and Spark to generate "Gold Standard" training datasets, directly supporting the R&D team in the precise alignment, evaluation, and fine-tuning of state-of-the-art foundation models.
- Refactored legacy data ingestion scripts to optimize cluster resource utilization, successfully reducing data processing latency by 20% and ensuring high-fidelity data availability for critical model alignment and downstream ML tasks.

KEY PROJECTS

HomieHub: Production MLOps Architecture **Link**

Tech Stack: Python, Google Cloud Run, Airflow, Docker, LangChain

- Deployed a production MLOps ecosystem on Google Cloud Run using Docker and Apache Airflow, creating a hybrid retrieval engine with a custom LLM Agent to autonomously sanitize, structure, and analyze high-volume unstructured social data streams.
- Enforced strict dataset reproducibility using DVC while automating deployment gates via GitHub Actions CI/CD pipelines, integrating Fairlearn to validate ranking parity and ensure unbiased compatibility scoring across live production environments.

PhysioPro: GenAI-Driven Motion Correction **Link**

Tech Stack: Snowflake Cortex, Mistral 7B, OpenCV, MediaPipe, Streamlit

- Engineered a low-latency 3D motion analysis pipeline utilizing OpenCV and Snowflake Cortex (Mistral 7B) to calculate pose deviations, generating instant, text-based corrective feedback overlays that directly improved patient exercise adherence by 25%.
- Optimized backend performance by structuring complex pose keypoints within Snowflake tables for rapid querying, and developed an interactive Streamlit dashboard to visualize recovery metrics, reducing manual clinical assessment time by 40%.

EDUCATION

Northeastern University, Boston, MA **December 2025**
Master of Science in Data Analytics Engineering *GPA: 3.8/4.0*

- **Courses**: Machine Learning Operations (MLOps), Gen AI with LLM in Data Engineering, LLM-based Dialogue Agents, NLP
- **Teaching Assistant**: Storytelling for Data, Course builder for Applied Generative AI