SAMHITA KOLLURI

(781) 947-7769 | samhita.kolluri@gmail.com | linkedin.com/in/samhita-kolluri | Medium Blog | GitHub | Hugging Face | Portfolio

EXECUTIVE SUMMARY

AI Research Engineer with 3+ years of experience specializing in Multimodal Generative AI, Agentic Workflows, and SLM optimization. Expert in architecting autonomous agentic workflows and fine-tuned LLMs on GCP, supported by a specialized background in building efficient vectorized data pipelines for high-throughput enterprise production environments.

SKILLS

Languages & Frameworks: Python, SQL, PyTorch, TensorFlow, LangChain, LlamaIndex, HuggingFace Transformers.
AI & MLOps: RAG, Fine-Tuning (LoRA/QLoRA), Model Quantization, Vector DBs, Docker, Kubernetes, MLflow, DVC.
Data Engineering: GCP, Databricks, Snowflake, Apache Airflow, dbt, PySpark, Kafka, Teradata, Informatica.

WORK EXPERIENCE

Humanitarians AI | Boston, MA **January 2025 - May 2025**
AI Research Engineer **Link**

- Fine-tuned Llama-3 via PEFT/LoRA to extract insights from domain-specific text, increasing reasoning accuracy by 20% on adversarial AGI benchmarks designed to identify model hallucinations.
- Developed a low-latency RAG pipeline using ChromaDB and custom re-ranking to reduce retrieval error rates by 35% for 1M+ multimodal document streams via vectorized tuning.
- Engineered an automated evaluation framework for agentic decision-traces to provide root-cause interpretability and defect analysis for complex, multi-step autonomous reasoning loops.
- Optimized inference for 7B parameter LLMs to achieve 15% compute reduction while maintaining high-fidelity output integrity and adherence to predefined safety standards.

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

- Re-architected data retrieval pipelines for systematic ML models, achieving a 99% reduction in execution latency through vectorized operations and high-concurrency query tuning.
- Developed automated audit frameworks for model training sets to mitigate bias and ensure data integrity, reducing production model drift incidents by 25% for Fortune 500 AI agents.
- Engineered domain-specific knowledge graphs and high-dimensional feature stores for predictive models, replacing legacy workflows with optimized, ML-ready data integrity layers.

AI Engineer **March 2021 - August 2022**

- Engineered high-throughput PySpark pipelines on Databricks to sanitize 10TB+ text/code corpora, ensuring high-fidelity data quality for Foundation Model pre-training and alignment.
- Refactored data ingestion layers for model alignment (RLHF), reducing processing latency by 20% and ensuring 99.9% data fidelity for complex autonomous agent research tasks.

KEY PROJECTS

N.E.X.U.S: Network for ElevenLabs X-call User Scheduling (15 Voice AI Agents) **Demo Link**
Tech Stack: *Python, FastAPI, Redis, ElevenLabs, GCP, Google Maps*

- Engineered a high-concurrency "Swarm" architecture to orchestrate 15+ parallel autonomous voice agents, utilizing Redis soft-locks and a custom state machine to maintain 100% state integrity.
- Developed an LLM-based evaluation layer to audit agent reasoning traces and tool-use, implementing adversarial prompting to filter hallucinations and ensure 98% factual grounding.

HomieHub: Distributed MLOps Architecture & CI/CD Evaluation **Code**
Tech Stack: *Python, Google Cloud Run, Airflow, Docker, LangChain*

- Built a containerized ecosystem on GCP with automated evaluation gates, maintaining sub-100ms latency during 10x spikes in concurrent agentic validation requests.
- Orchestrated automated ETL workflows using Apache Airflow and DVC for strict dataset versioning, ensuring systematic validation and prevention of regressions in production NLP models.

PhysioPro: Generative AI-Driven Motion Correction (Vision AI) **Code**
Tech Stack: *Snowflake, Streamlit, Optimization, Data Engineering, Mistral 7B, OpenCV*

- Optimized Snowflake storage for high-dimensional pose keypoints, improving query performance by 35% and ensuring data integrity for real-time motion analysis pipelines.
- Engineered a data validation framework to sanitize and version complex motion datasets, reducing processing overhead by 40% through model quantization and systematic quality checks.

RESEARCH EXPERIENCE

Patent & Paper: AI/IoT Integrated Approach for COVID-19 Prevention & Screening, IEEE 7th I2CT; Engineered and validated automated diagnostic models using high-fidelity medical datasets (2021–2022). **Link**
Paper: "Post-Quantum Cybersecurity Through Lattice-Based Cryptography," IJERT Vol. 14; (July 2025). **Link**

EDUCATION

Northeastern University, Boston, MA | *GPA: 3.8/4.0* | *M.S in Data Analytics Engineering* **December 2025**
Research Focus: MLOps, Generative AI with LLMs in Data Engineering, LLMs-based dialogue agents, NLP
Teaching Assistant: Storytelling for Data, Applied Generative AI (Course builder)