Gautham Kolluru

GitHub | LinkedIn | Stack Overflow | HackerRank

+1 240-639-9934 | gautham.kolluru@gmail.com | thegauthams.com

Professional Summary

AI Engineer with deep expertise in **Generative AI, agentic workflows, and data infrastructure**. Currently building **production-grade LLM agents** for Cisco telecom clients, achieving **50% latency reductions** and **80% cost savings** through token-optimization algorithms. Proven track record bridging applied ML (as an **Instructor for 500+ students**) and scalable engineering (Petabyte-scale pipelines). Strong in **Python, SQL (Snowflake), LangGraph, and Kubernetes**, focused on enterprise AI systems and data-driven analytics.

Technical Expertise

AI & GenAI: Large Language Models (LLMs), LangGraph, Agentic Workflows, RAG, Semantic Search, Prompt Engineering, Scikit-learn, TensorFlow, PyTorch, NLP (Word2Vec, TF-IDF). * **Languages:** Python (Advanced - Pydantic, Pytest), SQL (Snowflake), Java, Scala, Rust (Intermediate), Bash. * **Data Engineering:** Apache Spark, Airflow, Kafka, Snowflake, AWS (S3, Redshift, EMR, Glue), ETL/ELT Pipelines. * **Infrastructure & Tools:** Kubernetes, Docker, CI/CD, Git, Linux, Jira.

Professional Experience

Cisco | Data Engineer (Focus on AI/ML & Data) | 2021-Present

- **Cut LLM costs and latency by ~80%** by building a **token-aware, multi-agent orchestration** pipeline in **LangGraph** with **end-to-end validation** for reliable network diagnostics.
- **Built near real-time anomaly detection system** analyzing 10,000+ network devices using parallel processing, preventing $10M+ in potential network outages.
- **Designed fault-tolerant data pipeline** with 99.99% uptime protecting $10B+ network infrastructure through real-time threat detection.

84.51° | Senior Data Engineer | 2020-2021

- **Architected enterprise-scale ML feature store** processing 20B rows (5TB+) with 40+ transformation steps, reducing feature engineering time by 80% for data science teams
- **Led performance optimization initiative** achieving 40% latency improvement through distributed computing optimization (Spark tuning, data partitioning strategies) for customer analytics pipelines
- **Implemented real-time recommendation system data pipeline** supporting personalized marketing campaigns, contributing to $50M+ incremental revenue

Cisco | Data / Machine Learning Engineer | 2019-2020

- **Built real-time multi-cloud sync** between GCP and Snowflake with zero data loss, enabling ML training workflows.
- **Delivered incremental ETL pipelines** in Airflow/Dataflow, cutting data processing costs by 35%.

Global Data | Software Engineer | 2018

- **Built a distributed web-scraping platform** (Scrapy, Celery, Redis) processing 1K+ pages daily for ML training data.
- **Optimized extraction with proxy rotation + multiprocessing**, delivering 3x throughput and 99.9% reliability.

Invincible Software Solutions | Software Developer | 2017-2018

- **Led a 5-engineer team** delivering 12 microservices-based apps for US federal clients, hitting 98% quality and on-time delivery.
- **Built real-time Tableau/SSRS dashboards** to drive stakeholder decisions.

Royal Oman Police | Project / Systems Engineer | 2013-2015

- **Built a high-performance OLTP/OLAP system** for a nationwide speed-camera network, handling 500K+ daily transactions.

Mott MacDonald | Automation Engineer | 2009-2012

- **Automated PLC/SCADA control systems** for water treatment facilities, improving operational efficiency by 40%.
- **Led deployments of automated monitoring systems** for major infrastructure projects (Yas Island development, housing communities).

Education

B. Tech in Computer Science Engineering — Andhra University, 2009

Additional Achievements

- **Scale & Reliability:** Delivered ML/data platforms processing 100T+ daily transactions with high availability and real-time responsiveness.
- **Business Impact:** Shipped ML-driven solutions that improved operational efficiency and decision-making across enterprise clients.