Siddhant Rajhans

New York City, NY | +1 2012341064 | siddhantrajhans@gmail.com | LinkedIn | GitHub

Summary

ML Engineer with 1.5+ years building production AI systems processing 400M+ records/month. Specialized in agentic architectures, multimodal RAG, and LLM evaluation frameworks. Published researcher (IEEE, Springer) pursuing MS in ML at Stevens, seeking to advance frontier model evaluation and reasoning capabilities.

Technical Skills

Languages: Python, C/C++, CUDA, SQL
ML & Deep Learning: PyTorch, TensorFlow, Scikit-learn, XGBoost, LightGBM, Transformers, Hugging Face
LLM, Agents & Evaluation: LangGraph, RAG, Agentic Orchestration, Prompt Engineering, Multi-turn Conversationality, Evaluation Frameworks, Benchmark Design, Guardrails, Model Reasoning
Data & Search: Apache Spark, Kafka, FAISS, Postgres/Pgvector, Redis, SQLite, S3, OpenSearch
MLOps & Cloud: AWS (SageMaker, EKS, EMR, S3), Docker, Airflow, MLflow, W&B, Grafana
Research & Visualization: Jupyter, Pandas, Matplotlib, PowerBI, Gradio, Tableau, Zotero, Microsoft Office Suite

Education

Master of Science in Machine Learning 2025–2026(December)
Stevens Institute of Technology *Hoboken, NJ*

- **Relevant Coursework: Grade A** Natural Language Processing, Applied ML, Data Modeling, Data Acquisition, Modeling and Analysis: Big Data Analytics

Bachelor of Technology in Computer Science & Engineering
Swami Rama Himalayan University *Dehradun, India*

Experience

Machine Learning Engineer Oct 2023–Dec 2024
Stealth Healt-tech Company (NDA)

- Achieved **<65s p95 latency** processing **400M+ healthcare records/month** via scalable ML pipelines on **AWS** with **Airflow** and **Docker**.
- Improved clinical data retrieval accuracy by **30%** by applying advanced NLP and multimodal models to unstructured medical data.
- Increased user engagement by **25%** through **agentic AI systems** enabling adaptive, personalized interactions.

Software Development Engineer Intern Apr 2023–Sep 2023
Stealth Healt-tech Company (NDA)

- Boosted platform responsiveness by **35%** and user engagement by **20%** by integrating ML models into interactive healthcare platforms.
- Reduced average request latency by **40%** and increased throughput **3x** by optimizing backend services, APIs, and data pipelines.

Research Publications

CNN-Based Detection Mechanism for Deepfake Image – *IEEE ICCE 2025*

The Evolving Landscape of Cloud Computing: AI Integration, Threats, Challenges and Security Concerns – *Springer, ICRTC 2025*

Machine Learning and AI in Cybersecurity: Insights and Solutions – *Springer, ICSPN 2023*

Projects

Messaging-Based Data Pipeline *Kafka, PySpark, Docker, Airflow, Hive, NiFi*

- Architected an end-to-end streaming pipeline processing **400M+ daily events** with **65s latency** (95th percentile).
- Implemented **Kappa architecture** with event-time processing, watermarking, and fault tolerance sustaining **5K events/sec** throughput (12K peak).

Multimodal RAG for Scientific Literature *LangGraph, SciBERT, CLIP, TAPAS, OCR, FAISS*

- Built a **textual + multimodal RAG system** orchestrated for **multi-turn conversational queries** over 500K+ papers.
- Implemented **agent evaluation metrics** for grounding and reranking across **2.3M figures, 890K tables, 410K equations**.
- Optimized FAISS retrieval achieving **94% recall** at sub-**100ms latency**, scalable for production workloads.

Certificates

Training and Fine-tuning LLMs – **W&B** (2026) [Link]
Machine Learning – **Stanford Online, Coursera** (2024) [Link]
Generative AI with Diffusion Models – NVIDIA (2025) [Link]
Cloud Computing Diploma – IBM (2023) [Link]