NIKHIL JULURI

+1 (773)-310-8811 — nikhiljuluri4@gmail.com — njulu@uic.edu — Chicago, IL — LinkedIn — GitHub — LeetCode

Professional Summary

AI/ML Engineer with 4.5+ years of experience building **high-performance Python** and **machine learning systems** for **production environments**. Experienced in **large-scale data processing**, **private LLM deployment**, and **scalable MLOps pipelines** using **MLflow**, **AWS SageMaker**, and **Vertex AI**. **AWS Certified**, **finalist at the Microsoft HackWithChicago hackathon**, and **HackPrinceton Spring 2026 sponsor-track runner-up** for building **Lazarus**, an autonomous clinical R&D swarm.

Technical Skills

- **Programming:** Python, C++, SQL, Java, JavaScript/TypeScript.
- **Python Performance & Data Processing:** Memory Management, High-Performance Python, Multiprocessing, Multithreading, AsyncIO, Parallel Processing, Large File Processing, Multi-Gigabyte Flat File Processing, Batch Processing.
- **Deep Learning & LLMs:** PyTorch, Transformer Architectures, Attention Mechanisms, Fine-Tuning (LoRA/PEFT), Model Evaluation, Private LLM Hosting, GPT Models, Llama Models, Code Translation Automation.
- **LLM Systems & Optimization:** ONNX, ONNX Runtime, TensorRT, KV Caching, Dynamic Batching, Decoding Strategies, Latency Optimization, Throughput Optimization, GPU Memory Optimization, Observability (Prometheus, Grafana).
- **Generative AI:** Multi-Agent Systems, RAG, Vector Databases (FAISS, Pinecone), Model Context Protocol (MCP).
- **Backend & Distributed Systems:** FastAPI, REST APIs, Distributed Systems Design, Kafka, RabbitMQ, Caching (Redis).
- **Data & Storage:** MongoDB, PostgreSQL, MySQL, ETL Pipelines, Data Ingestion Workflows, Schema Validation.
- **MLOps & Deployment:** Docker, Kubernetes, MLflow, Kubeflow, Model Deployment, Experiment Tracking, Model Versioning.
- **Cloud & Infrastructure:** AWS (EC2 GPU Instances, S3, Lambda, SageMaker, Bedrock), Vertex AI, Scalable System Design.

Technical Experience

Graduate Research Assistant - University of Illinois Chicago, Chicago, IL

- **Developed memory-efficient Python pipelines** for ingesting and processing multi-gigabyte transaction-style and financial telemetry flat files common in enterprise and consulting-scale workflows; utilized PyArrow and multiprocessing libraries to achieve a **memory reduction of 35%** and an ingestion **throughput increase of 40%**.
- **Developed private LLM workflows** with GPT and Llama models for code translation and audit-style validation workflows; achieved an accuracy **increase of 87%** for anomaly detection and a **2x reduction** in debugging times.
- **Implemented MLOps pipelines** with MLflow for deployment on AWS SageMaker and exposure to Google Vertex AI for traceable and compliant model deployments aligned with enterprise system requirements; **achieved a 30% reduction** in release turnaround times.

Data Analyst - SodexoMagic, Chicago, IL *Sep 2024 – May 2026*

- Created interactive dashboards in **Power BI** and **Tableau** linked to **SQL** databases that displayed operational KPIs based on **5,000+ records**, resulting in a **40% reduction** in reporting time and quicker decision-making processes.
- Executed data preprocessing operations on structured **SQL** and **NoSQL (JSON)** data sources using **Python** and **Pandas** libraries, transforming unstructured operational data into structured datasets for repeated analysis purposes.
- Automated data processing operations through **Python** programming and **SQL job scheduling**, leading to the efficient generation of datasets and saving about **3–4 hours weekly** from manual reporting efforts.

Software Engineer II — Machine Learning Engineer - Deloitte, Hyderabad, India Sep 2022 – Jul 2024

- Built enterprise **GenAI advisory assistants** using RAG over financial statements and market APIs with LangChain and vector databases, reducing manual analysis time by **40%**. Developed evaluation pipelines measuring **answer relevance (0.87)**, **groundedness (0.91)**, and **hallucination rate (6%)** to ensure regulatory compliance and factual accuracy.
- Improved relevance of **investment product recommendations, portfolio suggestions, and retirement plan guidance** by **25%** by adapting transformer-based foundation models through **LoRA/QLoRA fine-tuning** and context-aware prompt engineering techniques, including **few-shot** and **role-based** prompting techniques, while optimizing token utilization and inference cost.
- Deployed **scalable GenAI services on AWS Bedrock and SageMaker** with provisioned throughput models, orchestrated through FastAPI microservices and integrated with Lambda, S3, and EC2. Implemented monitoring for latency, cost, and model performance to maintain **99.9% system uptime** and reduce operational overhead by **30%**.

Software Engineer I — Machine Learning Engineer - Deloitte, Hyderabad, India May 2022 – Aug 2022

- Built **credit risk prediction pipelines** using **ETL**, **feature engineering**, and **XGBoost / Random Forest** with **5-fold cross-validation** (AUC **0.86**), improving loan approval accuracy by **18%**. Developed **hybrid investment recommendation systems** using **collaborative filtering**, **XGBoost ranking**, and **Thompson Sampling**, increasing **CTR by 15%** (**Precision@K, NDCG**).
- Built a **fraud detection pipeline** using **Isolation Forest** and **Gradient Boosting**, reducing **false positives by 25%**. Deployed **Dockerized FastAPI APIs on AWS Lambda** for **real-time fraud monitoring**.

Projects

Lazarus | *FastAPI, React, PostgreSQL, Neo4j, Redis, WebSockets, OpenAI, Gemini, PubMed, openFDA*

- Architected a full-stack clinical AI platform that transforms failed drug assets into ranked repurposing hypotheses using a **FastAPI + React/Vite control plane**, PostgreSQL operational ledger, and Neo4j biomedical knowledge graph.
- Built a typed **9-agent LLM orchestration DAG** with **14 persisted reasoning steps per run**, generating auditable outputs across hypothesis generation, skeptical review, evidence curation, trial strategy, effort estimation, and impact scoring.
- Engineered real-time WebSocket streaming with polling fallback, enabling operators to monitor live agent traces, confidence scores, human-review escalations, portfolio rankings, and executive-ready PDF blueprint generation.

TrustLayer | *Python, Streamlit, LangChain, ChromaDB, BM25, Sentence Transformers, OpenAI API*
- Built a trust-aware research assistant for local research-paper corpora by engineering an end-to-end RAG pipeline that ingests PDFs, enriches metadata, chunks documents, and indexes **5,000+ evidence chunks** with Chroma vector search.
- Improved answer reliability by implementing **hybrid and corrective retrieval**, combining dense embeddings, BM25 sparse search, reciprocal rank fusion, and cross-encoder reranking, recovering stronger evidence in $\sim$**35% of weak first-pass retrievals**.
- Increased transparency and reduced unsupported responses by $\sim$**40%** by adding **verification-based abstention** and an interactive Streamlit dashboard that surfaces evidence, justification, and confidence metrics, returning "Insufficient evidence" when support was weak.

BugOrbit | *React, TypeScript, FastAPI, Neo4j, RocketRide*
- Designed and built a graph-powered incident intelligence platform that transforms raw production telemetry into structured incidents, root-cause analysis, blast-radius insights, and remediation guidance across **10+ distributed service dependencies**.
- Engineered a FastAPI and Neo4j backend pipeline to normalize noisy observability payloads, analyze trace failures, persist service dependencies as a live graph, and maintain real-time state across **active and resolved incidents with $<$200 ms ingestion latency**.
- Developed an interactive React and TypeScript dashboard for live incident monitoring, dependency-graph exploration, fix recording, and similar-incident retrieval, reducing mean investigation time by $\sim$**30%**.

GraphRAG for Multi-Hop Question Answering | *Python, PyTorch Geometric, Sentence Transformers, Streamlit, OpenAI API*
- Built an end-to-end GraphRAG system for HotpotQA-style multi-hop QA, indexing **10,000 examples** into **263,113 text chunks** with dense retrieval, hybrid graph construction, and an interactive Streamlit interface.
- Designed a hybrid graph-retrieval pipeline with query-aware GraphSAGE, dense–GNN fusion, and PCST-based evidence selection to improve multi-document reasoning beyond standard top-k semantic search across **3 retrieval modes**.
- Achieved **0.6733 exact match** and **0.7535 F1** on a **300-question** evaluation set using the Fusion retriever, outperforming the dense baseline by **1.7 EM points** and **2.3 F1 points** in downstream answer quality.

PulseGrid – Real-Time Graph Based Disaster Response Optimization System
Neo4j, Python, FastAPI, WebSockets, Graph Algorithms, Maps API, OpenAI
- Designed a **real-time graph-based decision-making system** on **Neo4j**, where **60+ nodes** and **150+ relationship edges** were modeled across responders, shelters, hospitals, volunteers, and roads; facilitated **multi-hop constraint satisfaction** with **sub-100ms WebSocket updates** and **automatic SOS prioritization** processes.
- Created a **multi-step process for dispatch and routing** using **Priority Queue sorting**, **constrained Dijkstra's shortest path algorithm**, **Yen's K shortest path algorithm (k=3)**, and **Hungarian Algorithm**, handling **10-15 simultaneous events** and decreasing the time taken for responder deployment to about **45-50%** while creating **3 alternative routes per event**.
- Executed **resource and volunteer scheduling** using **Gale-Shapley algorithm** for optimal matching and **Min Cost-Max Flow problem**, obtaining **100% optimal task assignment** for volunteers and **optimized supply chain management** across multiple shelters, besides providing **real-time route animations** and **ETA tracking (sub-1 second)** along with **real-time instructions generated using OpenAI API**.

High-Performance LLM Inference and Evaluation Framework | *Python, PyTorch, vLLM, CUDA*
- Built a high-performance LLM inference framework to evaluate and compare transformer models, improving **token generation throughput by $\sim$30–45%** and reducing **end-to-end latency by $\sim$25%** through optimized **dynamic batching**, **KV-cache reuse**, and efficient request scheduling.
- Designed benchmarking pipelines across **5+ model configurations** to analyze **latency distribution (p50/p95)**, **tokens-per-second throughput**, and **GPU memory utilization**, enabling systematic performance tuning and identification of inference bottlenecks under **concurrent workloads**.

Education

University of Illinois Chicago, Chicago, IL Aug 2024 – May 2026
Master of Science in Computer Science GPA: 3.8
Coursework: Cloud Computing, Algorithms, DBMS, Big Data Mining, Text Mining, ML on Graphs, Deep Learning with NLP
Chaitanya Bharathi Institute of Technology, Hyderabad, India Aug 2018 – Jun 2022
Bachelor of Engineering in Electronics and Communication

Certificates and Awards

AWS Certified AI Practitioner; AWS Certified Generative AI Developer – Professional; 5x Salesforce Certified; Deloitte SPOT Award