NIKHIL JULURI

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

Professional Summary

Software Engineer with 3.5+ years evolving from ML model development to enterprise GenAI systems, specializing in multimodal RAG, agentic workflows, PEFT fine-tuning, prompt engineering, inference optimization, and MLOps on AWS.

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

Technical Skills

- **Programming & Data:** Python, TypeScript, JavaScript, Java, SQL, NumPy, Pandas, SciPy, Plotly
- **Machine Learning:** Regression/classification (XGBoost, Random Forest), clustering (K-Means, DBSCAN), recommendation systems (collaborative + content-based), hyperparameter tuning, cross-validation, model evaluation
- **GenAI & NLP Foundations:** PyTorch, Hugging Face Transformers, TensorFlow, spaCy, NLTK, Gensim, Tokenizers, vector search (FAISS, Weaviate, Pinecone)
- **Systems & MLOps:** Docker, Kubernetes, MLflow, CI/CD, model monitoring, drift detection, Node.js, Spring Boot, REST, GraphQL, React, Next.js, PostgreSQL, MySQL, ETL, SOQL, AWS BedRock, AWS SageMaker

Technical Experience

Graduate Research Assistant - University of Illinois at Chicago, Chicago, IL Jun 2025 – Present
- Designed multimodal RAG systems combining transformer-based LLaMA models with Vision Transformers for text-image reasoning, integrating FAISS similarity indexing and Weaviate to retrieve structured and unstructured data, improving contextual response quality by **30%** across domain-specific research benchmarks and evaluation datasets.
- Enhanced model efficiency through LoRA fine-tuning and alignment strategies while optimizing inference via mixed precision, KV-caching, dynamic batching, parallelism, flashedAttention and controlled decoding; leveraged Optuna for hyperparameter tuning, balancing latency, throughput, and response relevance across large-scale research workloads.
- Deployed GPU-accelerated LLM pipelines on AWS using Docker and Kubernetes, implementing CUDA-based optimizations, memory management, token-level monitoring, and retrieval latency tracking to support scalable experimentation, reproducible pipelines, and real-time performance visibility across multimodal research systems and prototype deployments environments.

Software Engineer — Machine Learning Engineer - Deloitte, Hyderabad, India Sep 2022 – Jul 2024
- Built enterprise GenAI assistants using RAG over financial statements, retirement plans, and external market APIs, integrating LangChain with vector databases to automate advisor workflows, reducing manual analysis time by **40%**. Implemented response evaluation using answer relevance, groundedness, and hallucination checks to ensure compliance and factual accuracy.
- Enhanced personalization by adapting transformer-based foundation models via LoRA/QLoRA and supervised fine-tuning, applying few-shot, role prompting, and controlled decoding (temperature, top-k/top-p). Improved recommendation relevance by **25%** while monitoring token usage and optimizing prompt design to reduce inference costs.
- Deployed scalable GenAI services on AWS Bedrock and SageMaker with provisioned throughput models, orchestrated via FastAPI, Lambda, S3, and EC2. Integrated structured RAG pipelines with database and API sources, implementing cost monitoring, latency tracking, and model performance dashboards to maintain **99.9% uptime** and reduce operational overhead by **30%**.

Software Engineer Intern — Machine Learning Engineer - Deloitte, Hyderabad, India May 2022 – Aug 2022
- Built credit risk ML pipelines using ETL, feature engineering, and XGBoost/Random Forest with 5-fold CV (AUC 0.86), improving loan approval accuracy by **18%**; extended to hybrid investment recommendations with collaborative filtering, content-based models, XGBoost ranking, and Thompson Sampling, increasing CTR by **15%** (Precision@K, NDCG).
- Completed the platform with fraud detection using Isolation Forest and Gradient Boosting (ROC-AUC, recall), reducing false positives by **25%**; automated pipelines with scikit-learn and Docker, deploying FastAPI REST services on AWS Lambda for near real-time transaction monitoring.

Projects

Adaptive LLM Evaluation & Self-Optimizing Agent Framework Jan 2025 – Present
LangGraph, DSPy, vLLM, TruLens, Guardrails AI, FastAPI, Prometheus, Grafana, Docker, AWS
- Built an adaptive LLM orchestration framework using LangGraph and DSPy to dynamically route queries across multiple foundation models based on task complexity, enabling cost-aware model selection and improving response quality consistency by **27%** across benchmarked workloads and evaluation suites using task-specific prompt templates.
- Implemented self-refinement and guardrail pipelines with structured output validation, hallucination detection, and automated response scoring using TruLens, introducing iterative feedback loops that reduced factual error rates by **32%** while preserving end-to-end latency under production constraints with automated regression testing.
- Established LLMOps/MLOps pipelines with vLLM-served async FastAPI services, integrating Prometheus/Grafana for token, latency, and throughput monitoring, automated CI/CD with Docker, and dynamic temperature/top-p tuning, reducing inference cost by **24%** while maintaining production reliability across multi-node deployments.

Certificates and Awards

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