NIKHIL JULURI

+1 (773)-310-8811 — nikhiljuluri4@gmail.com — njulu@uic.edu — 821 South Laflin, Chicago, IL — LinkedIn — GitHub — LeetCode

Professional Summary

Software Engineer with 2+ years of experience specializing in AI, ML, and GenAI solutions for production systems. Skilled in PyTorch, LoRA/QLoRA fine-tuning, vLLM inference optimization, RAG pipelines, MLOps, Docker, Kubernetes, MLflow, and cloud deployment.

Education

University of Illinois Chicago, Chicago, IL August 2024 – May 2026 (expected)
Master of Science in Computer Science GPA: 3.8
Relevant Coursework: Cloud Computing, Computer Algorithms, Database Management Systems, Big Data Mining, Data Mining Text Mining

Chaitanya Bharathi Institute of Technology, Hyderabad, India August 2018 – June 2022
Bachelor of Engineering in Electronics and Communication

Technical Skills

- **Programming Languages:** Python, TypeScript, JavaScript, Java, SQL
- **AI/ML Frameworks:** PyTorch, Hugging Face Transformers, LangChain, LangGraph, scikit-learn
- **Data Science & Analysis:** NumPy, Pandas, Matplotlib, Seaborn, SciPy, Plotly
- **NLP Libraries:** NLTK, Gensim, spaCy, Transformers, Tokenizers
- **LLM & RAG Technologies:** RAG Pipelines, Vector Databases (FAISS, Weaviate), vLLM, LoRA/QLoRA Fine-tuning, RLHF, Ray Tune
- **MLOps & Deployment:** MLflow, Docker, Kubernetes, FastAPI, AWS (Lambda, EC2, S3, API Gateway), CI/CD
- **Frontend Technologies:** React.js, Next.js, LWC (Lightning Web Components), Component Architecture
- **Backend & APIs:** Node.js, RESTful Services, GraphQL, Spring Boot
- **Database Management:** PostgreSQL, MySQL, SQL Optimization, ETL Pipelines, SOQL, Data Loader

Technical Experience

Graduate Research Assistant - University of Illinois at Chicago, Chicago, Illinois June 2025 – Present

- Working on building RAG pipelines and LLM inference workflows using PyTorch and Hugging Face Transformers - mostly experimented with LLaMA and DistilBERT models. Set up vector databases like FAISS and Weaviate to handle the retrieval part. Tried out supervised fine-tuning and some RLHF-based alignment techniques on domain-specific datasets to make the inference faster and more memory-efficient. Spent a lot of time with Ray Tune doing hyperparameter tuning, playing around with learning rates, batch sizes, and attention parameters until we got better performance.
- Put together scalable AI/ML pipelines that covered the whole process - data ingestion, ETL, exploratory analysis, and feature engineering. Used Pandas, NumPy, and scikit-learn pretty heavily for all of this. Made sure we had good data quality throughout and kept everything version-controlled so the experiments could be reproduced easily. This was really important for making our ML work research-ready.
- Got hands-on experience packaging and deploying models with Docker and Kubernetes on AWS. Set up basic monitoring and logging for our experimental inference workflows so we could see how things were running in real-time. This made it a lot easier to catch issues early and iterate quickly during development.

Software Engineer — ML & LLM Inference, GenAI Systems - Deloitte, Hyderabad, India Sep 2022 – July 2024

- Built AI-powered financial account systems using **RAG pipelines, LLM inference, vector databases, LangChain, and LangGraph**, deploying **FastAPI microservices with Docker and Kubernetes on AWS (Lambda, EC2, S3)** for real-time queries across millions of accounts. Optimized payment workflows with **Node.js, TypeScript, and Apex**, reducing processing time by **60%**, helping a client with a $50,000 Flex Account achieve **$10,000 turnover growth**. Developed **React.js and LWC interfaces** that increased engagement by **50%** and portfolio growth by **40%**, enabling retirees to boost IRA and 401(k) contributions by **$15,000 annually**.
- Fine-tuned transformer models using **LoRA/QLoRA with MLflow experiment tracking and Optuna hyperparameter optimization**, cutting inference latency by **25%**. Integrated **vLLM for high-throughput serving** with optimized GPU utilization, quantization, and KV caching. Built **RESTful and GraphQL APIs via AWS API Gateway and FastAPI Lambda** that improved data accuracy by **40%** and reduced support tickets by **30%**, helping families track 529 plans and reach **$20,000 annual savings goals**.
- Led data migrations using **SOQL, Data Loader, and ETL pipelines with PyTorch, NumPy, Pandas, and scikit-learn** for ML validation, achieving **98% accuracy** while reducing manual work by **30%** and migrating **$5M in legacy plans**. Integrated **Java microservices (Spring Boot, Hibernate) with Python FastAPI ML endpoints**, delivering **40% faster responses** and **25% higher satisfaction**. Enabled real-time annuity assessment through AI dashboards, growing client AUM by **$50,000**.
- Established **MLOps with MLflow for model lifecycle management** and built **CI/CD pipelines for automated deployment and monitoring with FastAPI**. Implemented continuous retraining, logging, and rollback systems maintaining **99.9% uptime**, supporting millions of transactions with rapid experimentation capabilities.

Software Engineer Intern — AI/ML & LLM Systems - Deloitte, Hyderabad, India May 2022 – Aug 2022

- Conducted **exploratory data analysis on financial datasets using Pandas, NumPy, and Matplotlib** to uncover portfolio trends and anomalies. Built **Python ML pipelines with PyTorch and scikit-learn** for predictive analytics and recommendation tasks. Experimented with a small proof-of-concept **LLM prototypes using DistilBERT and LLaMA for small-scale RAG retrieval pipelines** that provided AI-powered insights into account workflows.
- Created **React.js and LWC dashboards** to visualize ML insights and portfolio recommendations for end users. Integrated **RESTful APIs with AWS Lambda and API Gateway** to connect ML pipelines with dashboards, enabling real-time financial data access and updates. Applied **LoRA-style fine-tuning on smaller transformer models** like DistilBERT and LLaMA to experiment with personalized recommendations.
- Implemented **ETL pipelines** to clean, preprocess, and transform financial datasets, achieving around **95% data accuracy**. Packaged experiments in **Docker containers** for reproducibility and shared cloud deployment across the team. Contributed to **MLOps-inspired workflows including data versioning, model tracking, and lightweight deployment pipelines** that helped establish best practices for experimentation.

Projects

Financial Portfolio Automation using RAG & LLMs *Python, PyTorch, LangChain, FAISS, LoRA, FastAPI, Docker, AWS, React.js*

- Developed an AI-powered system to streamline financial account management and portfolio rebalancing for clients, integrating LLMs and RAG pipelines for context-aware recommendations. Implemented a Retrieval-Augmented Generation pipeline using LangChain with FAISS vector database to retrieve financial statements and ETF/mutual fund data, enabling accurate recommendations in real-time.
- Integrated LoRA adapters on DistilBERT for specialized financial text understanding, achieving 95% accuracy in extracting actionable insights from unstructured financial documents. Used Pandas and NumPy for exploratory data analysis on financial datasets to identify patterns and preprocess data for model training.
- Deployed APIs using FastAPI and Docker for inference and connected real-time front-end dashboards built with React.js, reducing portfolio update processing time by 60% and improving client engagement by 50%. Leveraged AWS Lambda and API Gateway for scalable cloud deployment.

Large Language Model Fine-Tuning & Inference *PyTorch, Hugging Face, LoRA, vLLM, FastAPI, Docker, Kubernetes, AWS EC2/EKS*

- Fine-tuned an open-source LLaMA-8B quantized model on domain-specific financial datasets to generate high-quality financial insights, and deployed optimized inference for production use. Applied LoRA adapters on curated financial portfolios and regulatory documents, improving domain-specific text generation and question answering capabilities.
- Conducted hyperparameter tuning with learning rate scheduling and batch size optimization, along with RLHF-inspired reward modeling on supervised datasets to align model outputs with financial compliance and client requirements. Used Pandas and NumPy for data preprocessing and feature engineering.
- Deployed inference using vLLM on AWS EC2 GPU cluster, integrated with FastAPI, Docker, and Kubernetes, enabling sub-200ms token latency for concurrent requests. Set up auto-scaling GPU pods for high throughput and connected with LangChain for enhanced retrieval capabilities.

End-to-End MLOps / LLMOps Pipeline for Financial AI Systems *PyTorch, Kubeflow, Airflow, MLflow, vLLM, Triton, Docker, Kubernetes, AWS EC2/EKS*

- Built a full MLOps and LLMOps infrastructure to streamline training, deployment, monitoring, and versioning of ML and LLM models in a production financial environment. Developed a reproducible ML/LLM pipeline using Kubeflow and Airflow for data ingestion, preprocessing (EDA and ETL with Pandas and NumPy), training, and model versioning, supporting both traditional ML models and LLM adapters.
- Integrated MLflow for experiment tracking, model registry, and automated deployment to Kubernetes clusters with Docker containers, ensuring seamless rollback and monitoring of production models. Applied LoRA fine-tuning techniques on Hugging Face models for domain-specific adaptations.
- Orchestrated high-performance LLM inference with vLLM and Triton, connected to FastAPI endpoints, monitoring throughput, latency, and token consumption. Achieved 30% cost reduction compared to naive GPU deployments while maintaining high availability across AWS EC2/EKS infrastructure.

Certificates and Awards

- 5x Salesforce Certified: Certified AI Associate, Sharing and Visibility Architect, Platform App Builder, Administrator, Platform Developer 1.
- AI and Machine Learning Internship Certificate - National Instruments (NI) and Cognibot (June 2020): Completed intensive 4-week program covering Artificial Intelligence, Machine Learning, and Industrial IoT with hands-on training and projects
- Awarded a SPOT Award from Deloitte for outstanding contributions to the project.