

MOULIK SHAH

☎ +1 646 209 1462

✉ mps10088@nyu.edu

🌐 [Portfolio](#)

📄 [moulik-shah](#)

👤 [MoulikShah](#)

Education

New York University

September 2024 – May 2026

Masters in Computer Engineering

New York, NY

- Coursework: Machine Learning, Natural Language Processing, Deep Learning, MLOps, Computer Vision, Web Search Engine

Mumbai University

August 2019 – June 2023

Bachelor of Technology in Electronics and Telecommunication Engineering

Mumbai, India

- Coursework: Data Structures and Algorithms, Object Oriented Programming, Python, Mathematics, Database Management System, Computer Networks, Neural Networks, Big Data Analytics, Artificial Intelligence, Image Processing

Experience

Scale AI - Generative AI

September 2025 – December 2025

Generative AI Engineer Intern

San Francisco, CA

- Improved **LLM reasoning** on math and programming tasks through structured prompting, debugging model-generated Python/C++ code, and iterative feedback loops; strengthened factual accuracy using context-aware refinement.
- Built **agentic AI workflows** with **MCP-style tool-calling**, enabling models to autonomously retrieve context, validate code outputs, and self-correct using regex parsing, test harnesses, and action-observation evaluation patterns.

New York University - Volatility Laboratory (V-Lab)

January 2025 – December 2025

Machine Learning Research Assistant

New York, NY

- Developed an **LLM-powered sentiment trading system** using **FinBERT** and financial news data, increased risk signal accuracy by **15%** and improved integration with **SRISK** and **GARCH** models for market forecasting.
- Designed a scalable **backtesting framework** and **real-time volatility forecasting pipelines**, leveraging **GARCH-LSTM models** and **Elasticsearch**, reduced data retrieval time by **25%** and optimized predictive analytics for risk management.

Cornsprng - Software Engineering & DevOps

June 2025 – August 2025

Data Engineering Intern

New York, NY

- Built cloud-based ETL pipelines in **AWS** and **Azure** to ingest tens of millions of equity-price records daily. Engineered features like **moving averages**, **RSI**, **volatility** and enriched corporate-actions data, improving model performance by **60%**.
- Developed alerting systems with **CloudWatch** and **EventBridge**, and implemented **infrastructure-as-code (Terraform)** and **CI/CD workflows** to shrink deployment cycles from hours to minutes. Explored ML workflows with **SageMaker** to support downstream modeling.

MIKO - AI Research and Development

July 2023 – August 2024

Data Scientist

Mumbai, India

- Engineered a **recommendation engine** for skill-based games using **reinforcement learning** with **PyTorch**, employing **unsupervised clustering** techniques to segment users by engagement behavior. Conducted **A/B testing** to validate model efficacy, achieving a **25%** increase in user retention and reinforcing community safety through behavior-based recommendations.
- Led research and development of **multilingual NLP pipelines**, integrating **embedding models** and **Named Entity Recognition (NER)** for enhanced speech recognition. Refined pipeline through iterative experimentation with various embeddings, achieving a **30%** accuracy boost and supporting inclusive, language-diverse interactions.
- Enhanced **Open-Domain Question Answering system** with **Dense Passage Retrieval (DPR)** and **Retrieval-Augmented Generation (RAG)** models, achieving a **15%** reduction in unanswered queries through iterative model configuration and validation, supporting real-time NLP applications in **multilingual environments**.
- Integrated **LLMs** and **advanced embedding techniques** into chatbot functions, applying risk control and **anomaly detection** methods to analyze user behavior and uphold platform integrity by flagging suspicious interactions.
- Built a high-performing **song search engine** and **personalized recommendation system** using the iHeart Music dataset, using **feature detection** to analyze song attributes and elevate user experience in web-based applications, supporting platform safety by prioritizing user-friendly interactions.

Projects

Distributed Biometric Verification Pipeline with Scalable MLOps Deployment

May 2025

Python, PyTorch, MLflow, Triton, Docker, Kubernetes, Terraform

- Built a production-grade biometric verification pipeline using **ArcFace/iResNet-50** with distributed training (**DDP**, **AMP**) and experiment tracking in **MLflow**; automated reproducible workflows via **Docker**, **Kubernetes**, and **Terraform**.
- Deployed optimized **ONNX models** on **Triton Inference Server** with **OpenVINO**, achieving high-throughput, low-latency inference; implemented **evaluation workflows**, **canary rollouts**, and **drift-triggered retraining** to ensure fairness, reliability and scalability.

Cross-Domain Adaptation through Soft Prompt Tuning in Low-Parameter Language Models

November 2024

Python, Pandas, LoRA, LLMs, Linux, GitHub

- Optimized **cross-domain adaptation** for low-parameter **LLMs** using **soft prompt tuning** (e.g., prefix tuning) on **T5-large**, achieving over **90% accuracy** in sentiment analysis across multiple domains.
- Improved **few-shot learning** in resource-constrained models by integrating **LoRA** for efficient fine-tuning, reducing computational cost while maintaining high generalization capabilities.

Technical Skills

Languages: C, C++, C#, Java, Python, R, SQL, HTML, CSS, JavaScript

Frameworks: TensorFlow, PyTorch, OpenCV, YOLO, Django, Matplotlib, NumPy, Pandas, Flask, Scikit-learn, React.js, Node.js, AngularJS, Three.js, Next.js, Tailwind CSS, Spring Boot, LangChain, Hadoop

DevOps & Cloud: Docker, Jenkins, Spark, AWS, Jira, Postman, GCP, Snowfalke, Kubernetes, Airflow, Redshift

Databases: MySQL, PostgreSQL, SQLite, Firebase, MongoDB, Oracle

Developer Tools: Linux, Lucid, Figma, HuggingFace, Tableau, Power BI, Microsoft Excel