SUMANYU MUKU

New York, NY ◇ (551) 689-7368 ◇ sm9485@nyu.edu
LinkedIn ◇ Github ◇ Portfolio

SUMMARY

Machine Learning Engineer with 3+ years of experience shipping production AI systems across LLM applications, RAG, ASR, Document AI, ML observability, Anomaly Detection, and Computer Vision. Strong backend/MLOps foundation with Python, PyTorch, FastAPI, AWS, Kubernetes, Redis, PostgreSQL, and Neo4j; published researcher in model fairness and medical AI.

TECHNICAL SKILLS

AI/ML: LLMs, RAG, ASR, prompt engineering, LLM evaluation, rubric-based scoring, document understanding, anomaly detection, model monitoring, computer vision, fairness/bias mitigation
Frameworks/Tools: PyTorch, TensorFlow, OpenCV, LangChain, Prophet, OpenAI APIs, ElevenLabs, scikit-learn
Backend/MLOps: Python, C/C++, Java, FastAPI, AWS ECS, Docker, Kubernetes, Redis, Kafka, OpenTelemetry, Git
Data/Systems: PostgreSQL, MongoDB, Neo4j, Google Sheets API, React, Next.js, WebRTC

PROFESSIONAL EXPERIENCE

Amira Learning New York, NY
Machine Learning Engineer *Jan 2025 – Present*

- Built a real-time multimodal LLM application powering conversational AI avatars for middle-school history, integrating OpenAI, ElevenLabs, WebRTC, FastAPI, and Next.js; achieved sub-500ms response latency.
- Designed an AI document-understanding pipeline that extracts curriculum PDFs, maps content to six Amira skill taxonomies, validates structured outputs, and generates calendar-aware schedules across 180–200 school days.
- Developed human-in-the-loop review workflows that generate Google Sheets for CSM validation, enabling iterative correction of LLM extraction, skill mapping, and scheduling outputs.
- Built an automated reading-comprehension assessment service using ASR transcription and LLM-based rubric scoring to evaluate student retells for recall, sequencing, and comprehension quality.

Causely New York, NY
Software Engineer – ML Systems *Jun 2023 – Jan 2025*

- Developed and deployed a production ML microservice to learn Conditional Probability Distributions for causal graphs generated from Kubernetes topology scrapers.
- Engineered an ML monitoring system using regression analysis and constraint solving to evaluate the dynamic state and health of Kubernetes environments under changing load.
- Designed an edge anomaly-detection pipeline using Meta Prophet; tuned models to reduce false positives by 10x for latency and request-rate metrics.
- Built a graph-augmented RAG DevOps Copilot using Llama 3 7B and Neo4j service-topology retrieval to generate troubleshooting insights and root-cause hypotheses.

Apple Seattle, WA
Machine Learning Intern *May 2022 – Aug 2022*

- Designed the Siri Observability pipeline using ensemble ML models to detect regressions and slow-growing trends across builds, devices, locales, and product areas.
- Reduced regression-detection turnaround time by 24x by automating near-real-time SLO breach detection and Slack-based stakeholder notifications.

Indian Institute of Technology New Delhi, India
Machine Learning Engineer *Jun 2020 – Aug 2021*

- Developed an attention-based medical imaging model for COVID-19 chest X-ray assessment, improving precision from 65.9% to 81.9% and recall from 17.5% to 71.8%.
- Invented a data-curation technique for fairer object detection, reducing measured bias by 27% and improving detection performance by 9%.

PUBLICATIONS AND PATENTS

Does Data Repair Lead to Fair Models? WACV 2022
AI-Assisted Chest X-Ray Assessment for COVID-19 European Radiology 2021
Survey of Black-Box Adversarial Attacks on Computer Vision Models arXiv
Assuring Performance in a Computing Environment Using an ADG US Patent Pending

EDUCATION

New York University, MS in Computer Science New York, NY 2021 – 2023
Delhi Technological University, BTech in Computer Science New Delhi, India 2016 – 2020