SNEHA MAURYA

New York, NY | +1 (646) 578-0650 | sm5755@columbia.edu | linkedin.com/in/sneha | github.com/sneha1012 | Website

Education

Columbia University — New York, NY
Master of Science in Data Science — *Aug. 2024 – Expected Dec 2025*
Relevant Courses: Statistical Inference & Modelling, Applied ML, DL for NLP, Data Analysis & Visualization

SRM University — Chennai, India
Bachelor of Technology in Computer Science and Engineering — *Aug. 2020 – May 2024*
Awards: Merit Based Scholarship (2023)

Professional Experience

NXP Semiconductors — Austin, TX
Data Science Intern – Manufacturing Quality — *May 2025 – Present*
- Developing an **NLP pipeline** to classify semiconductor tool downtime causes from **EMS logs** using **topic modeling**, **fuzzy keyword matching**, and **GenAI prompt engineering** for scalable root cause detection across **400+ tools**
- Built a **modular ingestion pipeline** to standardize **8D reports**, **Kaizens**, and **Lessons Learned** documents across fabrication and design systems using **Python**, **Dataiku**, and **Teradata** for centralized, queryable metadata tracking

Columbia Business School — New York, NY
Graduate Research Assistant — *Jan. 2025 – May 2025*
- Fine-tuned a **multi-modal RAG system** for CSRD reports using **BGE-M3**, **GMM-based retrieval**, and **Qwen2.5-VL** to extract structured insights from OCR, tables, and diagrams
- Processed **400+ pages** each of sustainability pdf data with **85% retrieval accuracy** across diverse report layouts
- Built a **Streamlit tool** for regulatory query resolution and relevance visualization, reducing review time by **60%**

Metropolis Healthcare — Mumbai, India
Data Science Intern — *May 2024 – Aug. 2024*
- Developed a **GPT-4–powered NLP system** to convert clinical diagnostic text into simplified summaries, improving communication and readability across **10K+ patient records**
- Designed and deployed **end-to-end ETL and NLP pipelines** on **AWS** to process **1,000+ daily medical records**, ensuring **scalable**, **reliable**, and **privacy-conscious** operations

National University of Singapore — Singapore, Singapore
Deep Learning Researcher — *Jun. 2023 – Nov. 2023*
- Constructed a kitchen safety monitoring system using **YOLOv8** with **89% mAP**, applying **fairness-aware object detection** across **50+ real-world categories** in high-noise indoor environments
- Deployed an ML pipeline with **JSON integration** and async triggers, reducing real-time response latency by **20%**

Skills

Programming: Python, C/C++, SQL, R, Bash, JavaScript
ML/AI: LLMs, PyTorch, HuggingFace, Transformers, RL, GenAI, Topic Modeling, SKLearn, XGBoost, LightGBM, CV
MLOps: Docker, TVM, ONNX Runtime, Streamlit, Ollama, AWS, Airflow, CI/CD, REST APIs, Kubernetes, vLLM
Data Engineering: ETL, Spark, Postgres, Pgvector, OpenSearch, DVC, Tableau, Dataiku, GraphQL

Projects

LLM-RLX – Reinforcement-Learned Inference Router (PyTorch, ONNX, PPO, HuggingFace, Docker) *Jun. 2025*
- Engineered a reinforcement learning controller for **real-time routing** between quantized and full-precision LLM backends using latency, accuracy, and cache feedback as reward
- Integrated HuggingFace, ONNX Runtime, and TVM to enable **multi-backend execution**, achieving **3–6x speedups** with accuracy drop under 2%
- Planned extension for GPU–NIC scheduling aligned with SmartNIC runtime and DPU offload design

AutoKernel – CUDA Kernel Profiler + Optimizer (PyCUDA, Nsight, Streamlit, Docker) *May 2025*
- Developed a **profiler-guided system** to optimize CUDA kernels **(e.g., GEMM, conv2d)** using Bayesian search over launch configs to improve occupancy, throughput, and **memory access efficiency**
- Dockerized the profiling pipeline and launched a **Streamlit dashboard** with visual traces from **Nsight** Compute to visualize performance gains across baseline and optimized configs

Doc-Query Assistant (Python, LangChain, Streamlit, Chromadb, Ollama) *Jan. 2025*
- Constructed a **RAG-based AI** system to answer structured queries, processing 500+ academic, financial, and legal documents (PDF & Markdown) with 95% accuracy
- Extended app with local model support via **Ollama and latency-optimized vector** caching using ChromaDB, reducing inference time and **API overhead by 30%**

Leadership and Activities

- **Teaching Assistant II** at Columbia Business School, Led 20+ coding sessions and supported python & SQL grading
- **Board Member** at Graduate Society of Women Engineers, Organized 8+ STEM events for 200+ students.