

SNEHA MAURYA

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

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), Ranked 10th in 120+ teams for full-stack development, Hackathon (2022)

PROFESSIONAL EXPERIENCE

Columbia Business School

New York, NY

Graduate Research Assistant

Jan. 2025 – Present

- Developed a **RAG-based system** to automate sustainability report analysis, extracting **workforce** and **gender diversity** disclosures from 1000+ corporate reports, streamlining **compliance tracking** for businesses
- Fine-tuning **GraphRAG** and **Vision-Language Models (Qwen 2.5 VL, 72-B)** to enhance data structuring and extraction from unstructured reports, **improving retrieval efficiency** by 30% for EGS metrics

Metropolis Healthcare

Mumbai, India

Data Science Intern

May 2024 – Aug. 2024

- Built a medical report analysis system using **GPT-4** and **NLP**, generating patient-friendly summaries from complex healthcare diagnosis data for 10K+ monthly users, improving comprehension by 30%
- Designed and deployed ETL and ML pipelines for structured healthcare data, applying NLP techniques and ensuring scalability for 1,000+ daily medical records on **AWS Elastic Beanstalk**
- Collaborated with business stakeholders to develop automated KPI dashboards using Excel and Tableau, reducing weekly reporting time by 10 hours and **driving 15% higher app retention** through data-driven insights

National University of Singapore

Singapore, Singapore

Deep Learning Researcher

Jun. 2023 – Nov. 2023

- Led a team of 4 researchers to develop real-time kitchen safety monitoring system using **YOLOv8**, **achieving 89% Mean Average Precision (mAP)** accuracy and processing 50+ objects per frame
- Engineered an automated safety alert pipeline integrating ML predictions with JSON framework, reducing incident **response time by 20%** and improving team's monitoring efficiency by 35%

SKILLS

Languages: Python, R, SQL, C/C++, Bash, JavaScript

ML/AI: PyTorch, HuggingFace, Tran LightGBM, XGBoost, SKLearn, Computer Vision, NLP, LLMs

MLOps & Deployment: Docker, Kubernetes, AWS (Elastic Beanstalk, Athena, S3), Airflow, DeepSpeed, vLLM, GitHub

Data Engineering: ETL, Apache Spark, Postgres/Pgvector, OpenSearch, GraphQL, DVC

Data Analysis & Visualization: Tableau, Excel, Time Series Forecasting, Statistical Modeling, Data Storytelling

PROJECTS

Doc-Query Assistant — Python, LangChain, Streamlit, Chromadb, Ollama

Dec. 2024 – Jan. 2025

- Developed a **RAG-based AI** system to answer structured queries, processing 500+ academic, financial, and legal documents (PDF & Markdown) with 95% accuracy
- Built an **interactive Streamlit app** integrated with Ollama, reducing manual effort by 30% by **caching embeddings in ChromaDB** for faster querying

SEE-Q/A — Flask, VLM, Docker, OpenAI API, PyTorch, Git

Oct. 2024 – Nov. 2024

- Built a vision-language AI web application using PyTorch and Flask, **achieving 82.4% accuracy in Visual Question Answering (VQA)** to assist visually impaired users
- Implemented an **audio-based answer** generation system, delivering spoken responses from AI-generated answers
- Enhanced OpenAI's CLIP, a multimodal vision-language model, by integrating a custom linear layer with cross-entropy loss, improving image-text alignment through specialized answer vocabulary and **optimized data pipelines**

Reddit Engagement Predictor — GRU, Word2Vec, Prophet, TextBlob, K-Means

Aug. 2024 – Oct. 2024

- Developed a GRU neural network combining Word2Vec text embeddings and numerical engagement metrics, **achieving 91.45% accuracy** in subreddit post popularity predictions
- Implemented sentiment polarity analysis (TextBlob), LDA topic modeling, clustering, and Prophet time-series forecasting to **uncover content growth patterns**

LEADERSHIP AND ACTIVITIES

- Teaching Assistant II** at Columbia Business School – Facilitated course management to streamline operations. Led 20+ coding discussions, assisted with SQL grading, and mentored students on data analysis assignments
- Board Member** at Graduate Society of Women Engineers – Organized 5+ **STEM workshops** and career panels, engaging 200+ students and expanding outreach