Sneha Maurya

New York, NY | (646) 578-0650 | sm5755@columbia.edu | linkedin.com/in/snehamaurya10/ | github.com/sneha1012

Education

Columbia University — New York, NY
Master of Science in Data Science — *Aug. 2024 – Dec. 2025*
Relevant Coursework: Statistical Inference, Applied Machine Learning, Deep Learning, Natural Language Processing

Graduate Teaching Assistant: Data Analysis, Databases for Business Analytics, Python (graduate-level courses)

SRM University — Chennai, IN
Bachelor of Technology in Computer Science and Engineering (Merit-based Scholarship) — *Aug. 2020 – May 2024*
Relevant Coursework: Data Structures, Operating Systems, Cloud Computing, Artificial Intelligence, Databases

Experience

IBM — New York, NY
Data Scientist — *Sep. 2025 – Jan. 2026*
- Built a supervised code-edit dataset by mining **12K+ Git commits** and curating **850 Java edit pairs**; fine-tuned **Qwen3-Coder (30B)** with graph-augmented, structure-aware representations, improving edit accuracy by **12%**
- Designed an evaluation pipeline combining **AST edit distance**, semantic similarity, and LLM-based judgments; validated behavior on **60K+ multi-file edits** using **syntax parsing and compile checks**

NXP Semiconductors — Austin, TX
Data Science Intern — *May 2025 – Dec. 2025*
- Reduced unnecessary Failure Analysis for customer quality complaints by enabling **rule-base & XGBoost defect triage** across dies, integrating **iPAT inline defect screening**, layout-instance proximity, and ATE test coverage, yielding **$10,385 labor cost avoidance per CQC**
- Built an **Equipment Management root-cause analytics system** by clustering **10K+ unstructured EMS logs** across **400+ manufacturing tools** using **BERTopic**, saving engineers **6 hrs/week** via Power BI dashboards
- Standardized **R Shiny & SQL ETL pipelines** by converting **200+ lessons-learned records** into Teradata JSON tables, improving cross-site traceability

Columbia Business School — New York, NY
Graduate Research Assistant — *Jan. 2025 – May 2025*
- Developed a **multimodal RAG** system for **ESG compliance** analysis under the CSRD, integrating Qwen2.5-VL for visual reasoning, **RolmOCR** for structured extraction, and BGE-M3 embeddings with hierarchical retrieval
- Applied a **QLoRA adapter** on **Qwen2.5-VL-7B** to enhance reasoning over table captions, **achieving 85% retrieval precision** and enabling accurate report interpretation through a **Streamlit** interface

Metropolis Healthcare — Mumbai, IN
Data Science Intern — *May 2024 – Aug. 2024*
- Engineered **GPT-3.5-based clinical NLP system** integrating **RAG** for retrieval of similar historical cases, generating patient-friendly summaries across **10K+ medical reports** and improving patient engagement by **20%**
- Deployed **AWS pipeline (S3, Lambda, Bedrock)** processing **200+ reports/day** with HIPAA compliance
- Collaborated cross-functionally to build **SQL** & Tableau dashboards on 2M+ records to define product metrics, analyze acquisition and usage patterns, and **support product decisions**, reducing manual reporting by 12+ hours weekly.

Technical Skills

Programming Languages: Python, SQL, Rust, Java, Shell, R
Frameworks: PyTorch, TensorFlow, LangChain, FastAPI, Kafka, Flask, Ray, Spark
Developer Tools: AWS, GCP, Docker, Kubernetes, Jira, Git, Jenkins, Splunk, CI/CD, MongoDB
Libraries: NumPy, Pandas, Scikit-learn, Hugging Face, Matplotlib, NLTK, OpenCV, SHAP, A/B Testing

Projects

Chronos: Autonomous Email Agent | *Llama 3, DPO, FastAPI* — *Nov. 2025 – Dec. 2025*
- Fine-tuned a **Llama 3** email drafting model using **Direct Preference Optimization (DPO)** on **2.4K** preference pairs; improved draft acceptance from **61% to 87%**
- Designed a **production-safe inference pipeline** and deployed via **FastAPI** with a Chrome extension, reducing average response time by **70%** across concurrent conversations

Neural Code Search Engine | *Information Retrieval, Ranking* — *Sep. 2025 – Oct. 2025*
- Developed a scalable semantic code search system across **50+ repositories** and **2M+ indexed code snippets** using hybrid retrieval and learned re-ranking
- Improved top-3 retrieval accuracy by **28%** through bias-aware ranking and offline/user-driven evaluation on **500+ queries**, reducing code discovery time from **15 minutes to 30 seconds**