Shashank Pandey

s.pandey@nyu.edu | linkedin.com/in/spandey02 | (646)-372-2576

Education

New York University New York, NY
*Master of Science in Computer Science; **GPA: 4.0/4.0*** *Aug. 2024 - May 2026*

Birla Institute of Technology and Science, Pilani Hyderabad, India
Bachelor of Engineering in Computer Science, Minor in Data Science *Oct. 2020 - Jul. 2024*

Experience

Data Scientist Intern May 2025 – present
DynPro *San Jose, CA*

- Developing data analytics tools for the company's proprietary internal analytics platform.

Graduate Assistant Jan 2025 – May 2025
NYU Courant Institute of Mathematical Sciences *New York, NY*

- Served on the course team for a Natural Language Processing course, handling evaluation of coding assignments and exams.
- Led office hours to answer more than 90 students' queries, and mentored student groups by providing guidance, advice, and feedback for course projects.

Software Engineer Intern Jan 2024 – Jun 2024
Indian Institute of Science *Bangalore, India*

- Worked at the Center of Data for Public Good on implementing differential privacy for agricultural and medical use cases, coordinating with state governments and national bodies.
- Formulated and released a synthetic data generator for large low-dimensional datasets using Python, capturing the distribution of the original data while maintaining the privacy of all users.
- Developed a query engine to compute differentially private statistics and histograms from a dataset, deploying the solution to **over 50 metropolitan bodies**.

Data Engineer Intern May 2023 – Aug 2023
PayPal *Chennai, India*

- Worked in the Global Analytics and Data Sciences's Global Fraud Solutions team on building a model to flag suspicious phones used by consumers, preventing account takeovers during login events.
- Developed a rule-based approach that improved the existing process, leading to **21% more phones** being found. Designed and optimized a pipeline for the end-to-end deployment of the model, **reducing running time by 88%**, also ensuring data consistency and model explainability.
- Estimated to prevent losses of an average **$2 million/month** by effectively identifying suspicious activity.

Publications

LeGen: End-to-end Legal Information Extraction using Generative Models *EMNLP 2024*

- Developed LeGen, an end-to-end legal information extraction system leveraging large language models to capture complex discourse structure and semantics from legal text, significantly reducing error propagation.
- Achieved up to **32.2% improvement over SOTA benchmarks** and secured paper acceptance at the NLLP workshop held in EMNLP 2024 (Core A* conference).

Projects

Ed-Tech Search Engine | *Spark, FAISS, Python* **2025**

- Developed a scalable, **multimodal educational search engine** that enables students to retrieve and summarize relevant SAT-level question-answer pairs and YouTube videos, with text and video content.
- Engineered distributed data pipelines using Apache Spark and Spark NLP to process and embed **over 600,000 QA pairs and 20,000 video transcripts/frames**, leveraging BERT and CLIP models for semantic similarity and cross-modal retrieval, and **implemented FAISS** for real-time, high-dimensional vector search at scale.
- Integrated OpenAI APIs to summarize resources and generate answers, deploying it using **Streamlit and Ngrok**.

Analyzing Impact of Social Media on Subway Ridership | *MongoDB, QDrant, Python* **2025**

- Collected and analyzed over **36,000** NYC subway-related tweets across 2024 using VADER and RoBERTa sentiment models, correlating social media sentiment with daily subway ridership data to identify changes in ridership.
- Found that viral, negative social media discourse consistently led to citywide declines in subway ridership **across neighborhoods of different socio-economic levels**. Quantified and visualized these changes in ridership to study the disparate impact across boroughs.

Technical Skills

Languages: Python, Java, C/C++, SQL, MATLAB, HTML/CSS, Assembly, R, JavaScript
Frameworks: BigQuery, Hive, Hadoop, Jupyter, TensorFlow, MongoDB, Spark, SpringBoot, Django, React, Keras
Tools/Libraries: Docker, Tableau, NLTK, Pandas, NumPy, Matplotlib, SpaCy, SciKit, Atlassian Jira, QDrant