

## EDUCATION

<b>University of Pennsylvania (School of Engineering and Applied Science)</b> <i>Master of Science in Engineering (MSE) in Data Science; GPA 3.9/4.0</i> <b>Coursework:</b> Machine Learning, Deep Learning, Generative Modeling, NLP, Machine Perception, Databases, Computer Systems	<b>Philadelphia, PA</b> <b>Aug 2023 – May 2025</b>
<b>Kirori Mal College (KMC – University of Delhi)</b> <i>Bachelor of Science (BS) in Statistics; GPA: 8.11/10</i>	<b>Delhi, India</b> <b>July 2017 – Aug 2020</b>

## EXPERIENCE

**Research Intern, [Computational Social Listening Lab \(UPenn\)](#)** May 2024 – Present

### Project 1: (Misinformation)

- Worked on identifying misinformation on social media and whether it relates to health outcomes segmented by race.
- Extracted linguistic features from posts using [DLATK](#) and applied LDA for topic modeling and performed correlation analysis.
- Built NLP pipelines to detect health-related misinformation from **~20K social media** posts using RoBERTa and LDA topic modeling; improved **classification precision by 20%** through alignment-based entailment.
- Consolidated 10K+ survey responses from multiple platforms into a secure MySQL server and engineered features in Pandas.

### Project 2: (IH Risk Model)

- Developed a pipeline to predict Incisional Hernia (IH) risk from unstructured operative notes and structured EHR data in a surgical cohort of 10k+ patients.
- Engineered a few-shot GPT-based extraction pipeline to label operative features (e.g., incision, ostomy) from redacted notes **and reduced noisy extractions by 30% vs. BERT embeddings**.
- Scaled the pipeline for 10k+ notes using **parallelized batch inference** and memory-aware data chunking.

**Data Science Intern (Full-time), [Universal Media \(PA, USA\)](#)** May 2024 – Aug 2024

- Led the development of **3+ data pipelines** using Azure Data Factory (ADF), facilitating the seamless ingestion and transformation of diverse data sources into the Azure environment.
- Developed python scripts for data transformation, stored them in Blob storage and executed them via batch activity in ADF.
- Drove product insights by building **Mixed Media time series models in Azure Synapse**, analyzing marketing channel impacts on media diversity metrics. Built **Power BI** dashboards to deliver actionable insights for optimizing client strategies.
- Authored **5 stored procedures** in SQL, automating repetitive tasks and improving query performance by over 30%.

**Assistant Manager (Full-time), [IIFL Finance Ltd](#)** Apr 2022 – July 2023

- Analyzed ETL process failures and created **10+** paginated reports in SSIS to help the management track 1000+ branches.
- Optimized & migrated complex SQL queries from an obsolete database server that improved the **reporting services by ~40%**.
- Digital Adoption:** Led a product-focused initiative to identify digitally savvy customers by engineering features and building ADF pipelines to track campaign behavior. Trained and deployed a Random Forest model (**with a 90% accuracy**) in Azure ML Studio; **exposed it as a REST endpoint consumed by marketing campaigns**, driving digital **disbursal adoption by 50%**.

## SELECTED PROJECTS

- Diffusion Transformer (2025):** Implemented PatchVAE with convolutional encoders and patch-based decoding for fine-grained feature extraction. Trained a Diffusion Transformer to sample from the latent space of PatchVAE, achieving a **30% reduction** in FID score and **2x greater feature diversity** compared to VAE-generated samples. [\[Link\]](#)
- Multithreaded Image Processing (2025):** Engineered a parallelized box blur algorithm in C++ using POSIX threads, achieving a 2.8x speedup (3251 ms → 1165 ms) on 4 cores by optimizing memory access, leveraging shared-memory synchronization, and **partitioning workloads across non-overlapping thread-local regions**. [\[Link\]](#)
- FitBit (2024):** Engineered a Django health chatbot leveraging PostgreSQL for robust patient data management, featuring an LLM-agnostic architecture with seamless model switching via **Langchain that reduced overhead by 40%**. Optimized memory usage for long conversations through entity extraction, and **selectively triggered LLM responses** to reduce latency. [\[Link\]](#)
- Ride Duration Prediction (2024):** Developed a production-ready ML pipeline to predict NYC taxi ride durations, implementing **orchestration with Prefect, experiment tracking with MLflow, and automated hyperparameter tuning**. Designed modular, reproducible workflows to simulate real-world deployment scenarios. [\[Link\]](#)

## TECHNICAL SKILLS

**Programming Languages:** Python, C/C++, SQL, R programming, JavaScript

**Frameworks:** PyTorch, React, NodeJS, MongoDB, HTML/CSS, Neo4J, OpenCV, Apache Spark, MLOps, PySpark, Django, Flask

**Platforms & Tools:** AWS, MLflow, CI/CD pipelining, Kubernetes, Docker, Airflow, Azure Data Factory, Azure DevOps, A/B testing