

## EDUCATION

**University of Pennsylvania (School of Engineering and Applied Science)**

Philadelphia, PA

*Master of Science in Engineering (MSE) in Data Science; GPA 3.9/4.0*

Aug 2023 – May 2025

**Coursework:** Adv. Deep Learning (Diffusion models, LLMs, Multimodal architectures), Generative Modeling, Adv. Machine Perception, Machine Learning, Statistical/Probability models in Marketing, Database & Information Systems, Computer Systems

**Kirori Mal College (KMC – University of Delhi)**

Delhi, India

*Bachelor of Science (BS) in Statistics; GPA: 8.11/10*

July 2017 – Aug 2020

**Coursework:** Adv. Probability, Calculus, Linear algebra, Real Analysis, Statistical Inference, Time Series, Op. Research

## EXPERIENCE

Research Intern, **Computational Social Listening Lab (UPenn)**

May 2024 – Present

- **Research Goal:** Working on identifying misinformation on social media (FB posts, YouTube interactions) and whether or not it relates to health outcomes for people categorized on the basis of race (white/black) & urbanity (rural/urban).
- Developed a URL based misinformation-detection model that detected around 1k misinformed posts out of 1mm FB posts.
- Extracted linguistic features from posts using [DLATK](#) and applied LDA for topic modeling- computed Pearson's Correlation for these topic distributions with depression scores computed from user surveys to identify key linguistic markers.
- Implemented **semantic similarity and entailment analysis** using a pre-trained **RoBERTa** model to detect misinformation based on post alignment with trusted claims.
- Unified data from various online survey platforms in a secured server via MySQL and performed feature engineering in Pandas to prepare data for further downstream tasks.

Data Science Intern, **Universal Media (PA, USA)**

May 2024 – Aug 2024

- Architected Azure SQL Database solutions, encompassing DDL scripts to enhance data management and reporting solutions.
- 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.
- Authored **5 stored procedures** in SQL, automating repetitive tasks and improving query performance by over 30%.

Assistant Manager, **IIFL Finance Ltd**

Apr 2022 – July 2023

- Analyzed ETL process failures by familiarizing myself with Azure Data Factory and created **10+** paginated reports leveraging SQL Server Report Services functionalities to help the senior management track the business performance of 1000+ branches across the country.
- Optimized & migrated complex SQL queries from an obsolete database server that improved the **reporting services by ~40%**.
- **Digital Adoption:** Performed analysis to ascertain the features of the customers opting for digital loan disbursements. Built a Random Forest model on Azure ML to scope out potential customers and drive digital revenue with an **accuracy of 90%**.

## PUBLICATIONS (PREPRINTS)

- [1] [Enhancing Retrieval in QA Systems with Derived Feature Association](#) (**Under Review** at IEEE Cloud Summit)  
Keyush Shah, Abhishek Goyal\*, Isaac Wasserman\* [\[ArXiv\]](#) [\[Github\]](#)

## 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 compared to VAE-generated samples, demonstrating superior diversity and image quality. [\[Github\]](#)
- **Instance Segmentation (2024):** Developed an instance segmentation model using SOLO with ResNet-101 and FPN, achieving precise multi-scale segmentation without bounding boxes. Enhanced spatial accuracy with CoordConv and optimized mask prediction with Dice Loss for stable, high-performance results. [\[Github Link\]](#)
- **FitBit(2024):** Engineered a Django health chatbot with PostgreSQL for efficient patient data management, featuring an LLM-agnostic architecture for seamless model switching using Langchain. Optimized memory for long conversations and implemented advanced entity extraction to enhance medical context and automate request escalation. [\[Github Link\]](#)

## TECHNICAL SKILLS

**Programming & Skills:** Python, MySQL, C++, R, Git, React, NodeJS, MongoDB, MATLAB, JavaScript, HTML/CSS, A/B testing

**Software & Tools:** PyTorch, Eclipse, Jupyter, AWS Sagemaker, AWS Bedrock, SSMS, CUDA, LaTeX, Airflow, Azure Data Factory

## EXTRA-CURRICULAR and ACHIEVEMENTS

**Theatre:** Core Team Member (The Players). Lead Actor and auditioned for a brand advertisement and two short films.

**Community:** Collaborated with National Service Scheme (NSS), Delhi University & worked toward child empowerment.