Omer Tafveez

+1 (734) 596-6185 | omertaf@umich.edu | LinkedIn | GitHub

Education

University of Michigan, Ann Arbor — Michigan, USA
MS. Information Sciences - Big Data Analytics Track — *Aug. 2025 – May 2027*
Relevant Coursework: Foundations in LLM, Advanced Databases, Web Development, Applied Parallel Computing

Lahore University of Management Sciences — Lahore, Pakistan
BSc Economics-Mathematics, Computer Science Minor — *Aug. 2021 – May 2025*
Relevant Coursework: **Machine Learning**, **Generative AI**, **Deep Learning**, Principles and Techniques of Data Science, **Advanced Topics in Machine Learning**, Probability, Calculus I/II, Linear Algebra, Operational Research I, Econometrics I/II, Introduction to Quantum Computing, **Convex Optimization**, Advanced Programming
Note: Graduate Level courses are in bold

Research Experience

Decoupled Gradient Knowledge Distillation — Nov 2024 - Feb 2025

- **Role:** Primary Researcher at CITY@LUMS Lab Supervised by Dr. Muhammad Tahir (LUMS.)
- **Objective:** To enhance the performance of Decoupled Knowledge Distillation by explicitly adding gradient information (GIs) of logits of the target and non-target classes in the objective function, enabling informed Gradient Independence
- **Methodology:**
  - Introduced a novel reinforcement term in Decoupled Knowledge Distillation's loss function by **maximizing the mean-squared error between gradients of target-class (TCKD) and non-target-class (NCKD) losses** with respect to all logits, forcing them to extract *complementary and non-redundant* features.
  - The maximization of gradient MSE was empirically shown to **regularize overconfidence**, maintain minimal necessary target-class certainty for correctness, and convert uncorrelated gradient behavior into *purposeful feature specialization.*
  - Designed toy probability experiments to interpret loss dynamics across varying confidence/correctness scenarios, revealing balanced fidelity to teacher knowledge and prevention of excessive logit sharpening.
  - Demonstrated that this coupling accelerates convergence by promoting earlier neural collapse and tighter intra-class feature clustering, leading to stronger generalization.
  - Designed filter heatmaps, line-plots, gradient-weighted activation mapping, and t-SNE to visualise the impacts on the distillation.
- **Results:** Conducted experiments on CIFAR-100, TinyImagenet on ResNet, Vision Transformer, ShuffleNetV1.
  - The loss penalizes the student if its target probability is less confident than the teacher's target, but only when the student's prediction is incorrect → Improves performance.
  - The loss also penalizes the student if it does not match the teacher's prediction (even when the teacher's prediction is incorrect) → Improves fidelity.
  - The same holds for non-target probabilities, ensuring better generalization.
  - The similarities between intra-class features show improvement over decoupled knowledge distillation, indicating potential neural collapse at convergence, leading to performance gains.
  - Achieved state-of-the-art performance improvements, with an average **6% improvement on standard datasets** and **15% improvement in perturbed datasets**

Beyond Uniform Query Distribution: Key-Driven Grouped Query Attention — May 2024 – July 2024

- **Role:** Primary Researcher at CSaLT Lab Supervised by Dr. Agha Ali Raza (LUMS.)
- **Objective:** To develop an informed version of grouped query attention that groups queries dynamically using information from keys and queries.
- **Methodology:**
  - Developed *Dynamic Grouped Query Attention* (DGQA), assigning queries to keys based on key importance measured via norm magnitude at each iteration. To mitigate training volatility from consistent key assignments, we introduced a time-window framework that updates keys using changes in their cached norms from the window's start.
  - Performed conversion of MHA module of pre-trained VIT into GQA, which was further fine-tuned for 5 epochs for each different variant of GQA.
  - Evaluated all the variants on Vision Transformer of size Small, Base, and Large over CIFAR-10, CIFAR-100, Food-101, and Tiny Imagenet.
  - Implemented a Dynamic GQA variant using Exponential Moving Averages (EMA) to cache key norms, replacing the absolute start–end window difference with an EMA-based update. This yielded smoother training dynamics and reduced volatility.
- **Results:** Conducted experiments using GQA as the baseline for all the datasets.
  - Dynamic GQA outperforms GQA on VIT-Large by approximately **3**%.
  - The true benefit of the proposed models only comes forward on larger scales, i.e., with larger models and bigger datasets.

Teaching Experience

Teaching Assistant – CS 5302 (Foundations of Generative AI) January 2025 – May 2025
LUMS *Lahore, Pakistan*

- Facilitated learning for **120+ students** as a part-time TA, contributing 20 hours per week with Dr. Agha Ali Raza
- Prepared quizzes on model compression and quantization, and a programming assignment on small LLaMA.
- Assisted in checking quizzes, programming assignments, reading assignments, and the final exam.
- Held **4 hours of weekly office hours** to assist students with problems in assignments and quizzes.
- Delivered tutorials on Backpropagation, LSTM, Transformers, Parameter-Efficient Fine-Tuning, and Quantization.

Teaching Assistant – CS 535 (Machine Learning) August 2024 – December 2024
LUMS *Lahore, Pakistan*

- Liaison **between 250 students** as a part-time TA, contributing 20 hours weekly with Dr. Agha Ali Raza
- Prepared quizzes and programming assignments on Naive Bayes Classifier, Decision Trees, K-Means, and Transformers.
- Held **regular office hours twice a week** to assist students with quizzes, assignments, and contestations.

Teaching Assistant – CS 334 (Principles & Techniques of Data Science) January 2024 – May 2024
LUMS *Lahore, Pakistan*

- Liaison **between 70 students** as a full-time TA, contributing 40 hours weekly with Dr. Mobin Javed
- Prepared assignments on Exploratory Data Analysis, Linear and Logistic Regression, and Causal Graphical Computation.
- Held regular office hours to assist students with quizzes, assignments, and contestations.
- Prepared comprehensive tutorials on causality and final project topics ranging from Large Language Models to Economics.

Technical Projects

Transformers for Time-series: A Statistical View
PyTorch, Transformers, GluonTS, Statsmodels, SciPy, Seaborn, NumPy

- Diagnosed instability in Autoformer predictions by evaluating structural breaks, revealing limitations beyond the MASE metric.
- Analyzed forecast segments to identify clustered breaks, increased volatility, and spurious detections leading to misclassifications.
- Developed visual diagnostics (heatmaps, hourly line plots) to examine temporal break patterns across segments.
- Designed a statistical *Trust Score* using three OLS-based tests; forecasted series **averaged 0.37**, **failing 2 out of 3 reliability checks.**

Portfolio Website
Tailwind CSS, JavaScript, Github Pages, HTML/CSS

- Designed and developed a compelling personal portfolio using Tailwind CSS, showcasing expertise across machine learning, AI research, and real-time projects.
- Engineered a clean, responsive interface with fast performance and intuitive navigation using modern JavaScript and utility-first styling.
- Embedded personal branding with clear contact links, GitHub/LinkedIn integration, and a downloadable resume for easy access by recruiters or collaborators.

AI-Powered Crypto Investment Advisor
LangChain, OpenAI API, Binance API, React, Tailwind CSS, Python

- Developed an intelligent investment assistant with direct access to a personal Binance crypto wallet via Binance API, enabling portfolio-aware recommendations.
- Leveraged LangChain and OpenAI API tools to retrieve and synthesize real-time data from financial websites, news sources, and Twitter regarding user-specified cryptocurrencies.
- Integrated a Python Interpreter for function-calling to perform statistical analyses on historical price trends, social sentiment, and user-defined risk appetite.
- Implemented dual confidence interval calculations—estimating accuracy in approximating expected returns and in forecasting future price movement—validating the model's advisory outputs.
- Built a sleek, responsive ReactJS + Tailwind CSS frontend to display portfolio insights, real-time analytics, and investment suggestions in an interactive dashboard.

Content Moderation using LLMs - CS 535 Machine Learning Project
PyTorch, Matplotlib, NLTK, Transformers, NumPy

- Deployed a fine-tuned *LLaMA-2* model with QLoRA using Uber's Ludwig framework for toxic content classification, achieving an **F1 score of 0.81.**
- Pre-processed 223,500 text samples by normalizing, removing extraneous punctuation, fixing encoding errors, and filtering emojis to improve data quality.

Inflation Persistence & Structural Break Analysis in South Asia

R, STATA, Time-series Econometrics, ggplot

- Conducted a 36-year (1987–2022) econometric study of inflation dynamics across South East Asian Countries, benchmarked against U.S. trends to assess regional and global interdependencies.
- Applied Vector Autoregression (VAR) for interdependency modeling, Vector Error Correction Model (VECM) and Engle–Granger ECM for co-integration analysis, and ARIMA for inflation persistence estimation. Incorporated Bai–Perron structural break tests within VAR and ARIMA frameworks to segment inflation regimes.
- Performed Augmented Dickey–Fuller (ADF) and KPSS tests for stationarity, ARCH tests for heteroscedasticity, Johansen co-integration tests, Ljung–Box tests for residual independence, and SSR/BIC optimization for breakpoint selection.
- Identified significant structural breaks aligned with historical crises (e.g., 1998 Asian crisis, 2008 global financial crisis, 2022 inflation surge). Variance decomposition revealed Sri Lanka's inflation highly driven by external shocks, while Pakistan's was largely self-driven.

Mobility Pattern Mining: Explaining Urban Development - CS334 Data Science Project

NumPy, Pandas, Leaflet, Scikit-learn, Matplotlib

- Designed a Linear Regression model to approximate urban development in Beijing, achieving an R^2 **of 0.69** and **71% prediction accuracy.**
- Processed over 500,000 data points and segmented Beijing into 8 zones using KMeans clustering, visualizing population and development patterns with hexplots, scatterplots, and heatmaps.
- Integrated vehicle type, travel time, traffic metrics, and zone features as regressors to generate actionable city planning recommendations.

Industry Experience

Data Science Intern July 2023 – August 2023

Lipton Teas and Infusions — Python, GeoPandas, OR-Tools, Google Distance API *Karachi, Pakistan*

- Implemented a dynamic route scheduling system to replace fixed delivery paths between **7 warehouses and 256 distributors.**
- Collected and processed travel time data via the Google Distance API, building a distance matrix for the warehouse–distributor pairs.
- Performed exploratory data analysis on historical delivery logs to identify bottlenecks, such as idle time and vehicle underutilization.
- Developed a Python-based workflow (Google OR-Tools) to assign deliveries to vehicles while respecting capacity, time window, pallet size, and item-type constraints.
- **Reduced average delivery time by ∼18%, cut transportation costs by ∼22%, and improved on-time delivery rate from 82% to 95%** compared to the fixed-route baseline.

Data Science Intern June 2022 – July 2022

Indus Motors Company (Toyota, Pakistan) — Python, Scikit-learn, Pandas, NumPy *Karachi, Pakistan*

- **Achieved 87.3% accuracy** using Ridge Regression to model the impact of T-Bills, inflation, and SBP policy rates on prices.
- Pinpointed key sales drivers with Decision Trees, helping address the underperformance of 5 key vehicles.
- Ensured time-series stationarity via EMA and differencing, validated through ACF plots and ADF testing, improving R^2 to 0.93.

Technical Skills

Languages, Libraries, and Frameworks: C++, JavaScript, TypeScript, SQL, Python, HTML/CSS, Pytorch, NumPy, SciPy, Statsmodel, Pandas, Tailwind CSS, ReactJS, NodeJS, OracleDB, MongoDB, Git, Bash, STATA, R, Prophet, Matplotlib, Seaborn, ggplot, Neuronpedia, SAELens