Md. Shariar Kabir

shariar1405076@gmail.com shariar076.github.io Shariar076 Google Scholar

Research Interests

Mechanistic interpretability of large language models, NLP, and data science for social science and psychometric modeling. Currently I am working on,

- Methods for understanding and controlling LLM socio-political reasoning and belief formation.
- Combining interpretability techniques with quantitative social science evaluation.
- Collective truth formation in LLMs to reduce out-of-distribution behaviors.

Publications

Beyond the Surface: Probing the Ideological Depth of Large Language Models. **In Progress**
***Shariar Kabir**, Kevin Esterling, and Yue Dong.* *arXiv*

Our findings suggest that ideological depth is a quantifiable property of LLMs and that steerability serves as a valuable window into their latent political architecture.

Testing Conviction: An Argumentative Framework for Measuring LLM Political Stability. **In Progress**
***Shariar Kabir**, Kevin Esterling, and Yue Dong.* *arXiv*

Our findings demonstrate that ideological stability is topic-dependent and challenge the notion of monolithic LLM ideologies, and offer a robust way to distinguish genuine alignment from performative behavior.

Automatic Speech Recognition for Biomedical Data in Bengali Language. **2024**
***Shariar Kabir**, Nazmun Nahar, Shyamasree Saha, and Mamunur Rashid.* *arXiv*

This paper presents a prototype speech recognition system specifically designed for Bengali biomedical data.

SynthNID: Synthetic Data to Improve End-to-end Bangla Document Key Information Extraction. **2023**
*Syed Mostofa Monsur, **Shariar Kabir**, and Sakib Chowdhury.* *EMNLP Workshop*

We introduce SynthNID, a system to generate domain-specific document image data for training OCR-less end-to-end Key Information Extraction systems.

Education

Bangladesh University of Engineering and Technology (BUET) **2019 - 2022**
M.Sc. in Computer Science & Engineering (Part-time, coursework completed) *CGPA: 3.54/4.00*

Thesis: Dynamic Resource Allocation for Workloads in Serverless Architecture – predictive modeling using collaborative filtering & SVD for optimal cloud resource configuration. [paper]

Coursework: Bioinformatics Algorithms, Distributed Computing Systems, Data Mining, Data Management in the Cloud, etc.

Bangladesh University of Engineering and Technology (BUET) **2015 - 2019**
BSc in Computer Science & Engineering *CGPA: 3.53/4.00*

Thesis: Active Learning on Big Data – applied active learning to distributed datasets for scalable classification. [report] [ppt]

Selected Coursework: Machine Learning, Pattern Recognition, Artificial Intelligence, Digital Image Processing, etc.

Research Experience

University of California, Riverside **Summer 2025 – Present**
Research Intern – Fall'26 PhD aspirant, NLP Lab (Prof. Yue Dong) *Riverside, CA*

Mechanistic Interpretability of LLM in Socio-Political Reasoning

- Analyzing activation pathways linked to ideological responses using SAE features from Neuronpedia.
- Evaluating steerability and belief formation mechanisms in LLMs.

Analysing LLMs' Social Epistemology using Bayesian Statistics

- Implemented Multidimensional IRT from scratch in Stan; validated with 98% correlation to DW-NOMINATE scores.
- Applied to model LLM ideological positioning from questionnaire responses.

Industry Research Experience

Celloscope Ltd. **2020 – Present**
Senior AI Research Engineer *Dhaka, Bangladesh*

- Designed Bengali Conversational AI systems using **BERT-based NLU** and **retrieval-augmented generation**.
- Led monitoring human behavior research for various domain-specific applications.

MedAI Pvt. Ltd. **2021 – 2024**
Data Scientist (Part-time) *Cambridge, UK*

- Developed Conversational AI for mental health symptoms, integrating symptom classification models.
- Built multimodal disease prediction pipelines and fine-tuned **Whisper ASR** incorporating structured & unstructured clinical data.

Skills

Research:	Mechanistic Interpretability, Topic Modeling, Item Response Theory, Model Visualization
Programming:	Python, Shell, C, C++, STAN, LaTeX, SQL, TypeQL
Machine Learning:	PyTorch, ScikitLearn, OpenCV, Pandas, Datasets, Transformers, SpaCy
Tools:	LangChain, Neuronpedia, SGLang, Ollama, OpenAI, Spark, PySpark, Docker
Soft-Skills:	Communication, Collaboration, Presentation, Technical Writing

Selected Projects

Exercise Monitoring System: Inspired by research works like VidDiff and HuMMan. We created a system for LG Nova's Real-Time AI Fitness Coaching. Our system leverages Vision-Language Models (VLMs) to assist users in performing exercises correctly by comparing their execution against reference videos of expert demonstrations.

Bengali Conversational AI: Built with Bengali ASR, fine-tuned NLU engine, and dynamic conversational AI for natural language banking interactions. Implemented **vector index search** over FAQ datasets and integrated speech-to-text and text-to-speech pipelines for seamless voice-driven customer service.

ASR System for Patient Symptoms: ASR system for understanding medical symptoms spoken by patients in the Bengali language. I performed a comparative analysis using **DeepSpeech** and **Whisper** (tiny) model, using audio data collected from consented users using an audio data collection portal I designed.

End-to-end Document Key Information Extraction Using Synthetic Data: In this work, we propose a simple synthetic document image generation framework for Bengali documents. We show that the generated data improves the performance of the extraction model on real datasets. The short paper was accepted in the BLP Workshop at EMNLP'23.

wQFMSpark – Performance Analysis of Species Tree Estimation Using wQFM in a Distributed System: Species tree estimation from Gene Trees is crucial in *Phylogeny*. Quartet-based estimation techniques like ASTRAL, QMC, and QFM are widely used, but some struggle with scalability. This project aims to redesign wQFM for scalability. [report] [code]

Awards & Achievements

Industry Coding Assessment **2025**
CodeSignal General Coding Assessment (ICA): 510/600 (≈ 722/850 equivalent GCA, top 15%)

Global Health Equity Challenge Award **2024**
MIT Solve
Recognized for innovative approach to accessible healthcare (Top 6/2200+). [link]

References

Prof. Yue Dong **yue.dong@ucr.edu**
Assistant Professor in Computer Science and Engineering, University of California, Riverside

Prof. Kevin Esterling **kevin.esterling@ucr.edu**
Professor of Public Policy and Political Science, University of California, Riverside

Prof. Muhammad Abdullah Adnan **adnan@cse.buet.ac.bd**
Professor in CSE, Bangladesh University of Engineering and Technology (BUET)