Md. Shariar Kabir

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

Research Interests

Mechanistic interpretability and psychometric modeling of large language models with a focus on their behavior in longer context and multi-turn conversations.
Currently I am working on,

- Methods for understanding and controlling LLM socio-political reasoning and response stability.
- Combining interpretability with intervention finetuning for controllable knowledge-depth formation.
- Quantitative social science evaluation of LLMs to evaluate out-of-distribution behaviors.

Publications

[4] Shariar Kabir, Kevin Esterling, and Yue Dong. PReSS: A Black-Box Framework for Evaluating Political Stance Stability in LLMs via Argumentative Pressure. *arXiv preprint arXiv:2504.17052*, 2025. Under Review.

[3] Shariar Kabir, Kevin Esterling, and Yue Dong. Beyond the surface: Probing the ideological depth of large language models. *arXiv preprint arXiv:2508.21448*, 2025. In Progress.

[2] Shariar Kabir, Nazmun Nahar, Shyamasree Saha, and Mamunur Rashid. Automatic speech recognition for biomedical data in Bengali language. *arXiv preprint arXiv:2406.12931*, 2024.

[1] Syed Mostofa Monsur, Shariar Kabir, and Sakib Chowdhury. Synthnid: Synthetic data to improve end-to-end Bangla document key information extraction. In *EMNLP 2023 Workshop on Bangla Language Processing*, 2023.

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*

Understanding LLMs' response instability over longer context.

- Finding the correlation between instability and epistemic uncertainty over multiturn conversation.
- Evaluating stability after model finetuning and activation steering.

Mechanistic Interpretability of LLM in Socio-Political Reasoning

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

LLMs' Social Epistemology using Bayesian Statistics

- Implementing Multidimensional IRT from scratch in Stan (validated with 98% correlation to DW-NOMINATE scores).
- Applying psychometric methods to model LLM ideological positioning compared to humans.

Industry Research Experience

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

- Led numerous NLP research-based projects involving video monitoring and object classification.
- Engineered private and self-hosted Conversational AI solutions using open-sourced LLMs and RAG.

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

- Collected and curated Bengali biomedical audio data for fine-tuning Whisper.
- Built multimodal disease prediction pipelines 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)