Md. Shariar Kabir

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

Research Interests

Mechanistic interpretability and psychometric modeling of large foundation 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.

Selected 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.

**** indicates co-first author.***

Research Experience

University of California, Riverside **Winter 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 conversations.
- 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 a number of NLP and computer vision-based projects deployed across multiple industrial domains.
- 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.
- Evaluated Conversational AI for mental health symptoms, integrating symptom classification models.

Education

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

Thesis: Dynamic Resource Allocation for Workloads in Serverless Architecture. [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 for scalable classification usinf distributed infrastructure. [dissertation]
Selected Coursework: Machine Learning, Pattern Recognition, Artificial Intelligence, Digital Image Processing, etc.

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

Medical Classification by Probing LLMs: Multi-label classification of medical disciplines by training linear probes on activations from LLMs pretrained on Medical data. We extract layer-wise attention head activations from medical-domain LLMs and use Ridge regression classifiers to predict relevant medical disciplines from clinical descriptions. [code]

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.

Drawing Checker: An initiative to automate the design-error detection and verification in engineering drawings. The system used computer vision techniques and generative models to evaluate technical drawings, identify inconsistencies, and flag deviations from design constraints. I directed the model training and dataset curation pipelines, ensuring that the models achieved consistent accuracy across diverse geometric and structural inputs.

Resume Shortlister: An NLP-driven retrieval and ranking system designed to automate candidate selection for enterprise recruitment. By designing a hybrid RAG approach combining rule-based filtering with semantic retrieval, we developed a system capable of aligning candidate attributes with organizational requirements.

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