Abhik Bhattacharjee

38/9/A, Mirpur-11, Dhaka, Bangladesh
+880-1954689352

abhik@ra.cse.buet.ac.bd
https://abhik.vercel.app/
Google Scholar | GitHub

Research Interests

- Data and Computing-Efficient Deep Learning
- Robust and Adaptive Evaluation for NLP Systems
- Equitable AI Safety and Trustworthy Systems
- Low-Resource, Multilingual, and Cross-Lingual NLP

Education

- **Bangladesh University of Engineering and Technology (BUET)** — Dhaka, Bangladesh
  B.Sc. in Computer Science and Engineering — Feb 2016 - Feb 2021
  CGPA: 3.69/4 (Major GPA: 3.81) | Rank: 26^{th}/143

Publications

(* indicates equal contribution) | **h-index: 8 (Citations: 1,100+)**

1. **CrossSum: Beyond English-Centric Cross-Lingual Summarization for 1,500+ Language Pairs**
   Abhik Bhattacharjee*, Tahmid Hasan*, Wasi U. Ahmad, Yuan-Fang Li, Yong-Bin Kang, Rifat Shahriyar
   Proceedings of 61st Annual Meeting of the Association for Computational Linguistics: ACL 2023. [PDF] [Code]

2. **XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages**
   Tahmid Hasan*, **Abhik Bhattacharjee***, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, Rifat Shahriyar
   Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. [PDF] [Code]

3. **BanglaBERT: Language Model Pretraining and Evaluation Benchmarks for Low-Resource Language Understanding Evaluation in Bangla**
   Abhik Bhattacharjee*, Tahmid Hasan*, Wasi Uddin Ahmad, Kazi Samin, Md Saiful Islam, M. Sohel Rahman, Anindya Iqbal, Rifat Shahriyar
   Findings of the North American Chapter of the Association for Computational Linguistics: NAACL 2022. [PDF] [Code]

4. **BanglaNLG: Benchmarks and Resources for Evaluating Low-Resource Natural Language Generation in Bangla**
   Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Rifat Shahriyar
   Findings of the Association for Computational Linguistics: EACL 2023. [PDF] [Code]

5. **Not Low-Resource Anymore: Aligner Ensembling, Batch Filtering, and New Datasets for Bengali-English Machine Translation**
   Tahmid Hasan*, **Abhik Bhattacharjee***, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, Rifat Shahriyar
   Proceedings of the Empirical Methods in Natural Language Processing, EMNLP 2020. [PDF] [Code]

6. **IllusionVQA: A Challenging Optical Illusion Dataset for Vision Language Models**
   Haz Sameen Shahgir*, Khondker Salman Sayeed*, **Abhik Bhattacharjee**, Wasi U. Ahmad, Yue Dong, Rifat Shahriyar
   Proceedings of the First Conference On Language Modeling: COLM 2024. [PDF] [Code]

7. **Characteristics of Bias upon Context Length Variations: An Empirical Study on Bangla**
   Jayanta Sadhu*, Ayan Antik Khan*, **Abhik Bhattacharjee**, Rifat Shahriyar
   Findings of the Association for Computational Linguistics: ACL 2024. [PDF] [Code]

8. **GEMv2: Multilingual NLG Benchmarking in a Single Line of Code**
   Sebastian Gehrmann, **Abhik Bhattacharjee**, ···
   Proceedings of the Empirical Methods in Natural Language Processing, EMNLP 2022. [PDF] [Code]

9. **BanglaParaphrase: A High-Quality Bangla Paraphrase Dataset**
   Ajwad Akil*, Najrin Sultana*, **Abhik Bhattacharjee**, Rifat Shahriyar
   Proceedings of the Asia-Pacific Chapter of the Association for Computational Linguistics: AACL 2022. [PDF] [Code]

Preprints:

1. **Multi-ToM: Evaluating Multilingual Theory of Mind Capabilities in Large Language Models**
   Jayanta Sadhu, Ayan Antik Khan, Noshin Nawal, Sanju Basak, **Abhik Bhattacharjee**, Rifat Shahriyar
   ArXiv Pre-print, 2024. [PDF] [Code]

Selected Research Projects

1. **CrossSum: Beyond English-Centric Cross-Lingual Summarization for 1,500+ Language Pairs**

   Supervisors: *Prof. Rifat Shahriyar (BUET)* and *Dr. Wasi Uddin Ahmad (NVIDIA AI)*
   Status: Published in *ACL, 2023*

   - Curated *CrossSum*, a large-scale dataset with 1.7 million article-summary samples across 1500+ language pairs by aligning identical articles via cross-lingual retrieval.
   - Developed a novel multistage data sampling algorithm to effectively train models with explicit cross-lingual signals, enabling summary generation in any target language.
   - Proposed *LaSE*, a novel metric for automatic evaluation of cross-lingual summaries when target-language references are unavailable.

2. **XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages**

   Supervisors: *Prof. Rifat Shahriyar (BUET)* and *Dr. Yuan-Fang Li (Monash Uni.)*
   Status: Published in *Findings of ACL, 2021.*

   - Presented *XL-Sum*, a comprehensive dataset of 1 million professionally annotated article-summary pairs in 44 languages, curated from BBC News using robust extraction heuristics.
   - Established strong multilingual benchmark results (>11 ROUGE-2 on ten languages) to demonstrate the dataset's efficacy for training multilingual summarization models.

3. **BanglaBERT: Language Model Pretraining and Evaluation Benchmarks for Low-Resource Language Understanding Evaluation in Bangla**

   Supervisors: *Prof. Rifat Shahriyar (BUET)* and *Dr. Wasi Uddin Ahmad (NVIDIA AI)*
   Status: Published in *Findings of NAACL 2022*

   - Developed *BanglaBERT*, a state-of-the-art BERT-based model for Bangla, pre-trained on a meticulously curated corpus with 2B+ tokens.

- Established the *Bangla Language Understanding Benchmark (BLUB)* and introduced two new downstream datasets for comprehensive evaluation.
- Built *BanglishBERT*, a model jointly trained on Bangla and English, facilitating strong zero-shot cross-lingual transfer performance.

4. **Not Low-Resource Anymore: Aligner Ensembling, Batch Filtering, and New Datasets for Bengali-English Machine Translation**

   Supervisors: *Prof. Rifat Shahriyar (BUET)* and *Prof. M. Sohel Rahman (BUET)*
   Status: Published in *EMNLP, 2020.*

   - Built a customized sentence segmenter for Bengali to address issues of erroneous segmentation and noise in translation corpora.
   - Introduced two novel methods for sentence alignment from noisy comparable corpora on low-resource setups: *Aligner Ensembling* and *Batch Filtering.*

5. **Improving Document-Level Event Argument Extraction with Coreference resolution**

   Supervisor: *Prof. Rifat Shahriyar (BUET)*
   Status: Completed

   - Demonstrated the efficacy of explicit coreference resolution within the conditional generation framework for document-level event argument extraction.
   - Employed a contrastive learning loss formulation among entity mentions from the same coreference cluster to improve end-to-end argument identification performance.

6. **S3C: Program Synthesis using Self-Sampling and Self-Correction with Execution Feedback**

   Supervisor: *Dr. Abdus Salam Azad (WandAI)*
   Status: Completed

   - Proposed an iterative training paradigm (S3C), where CodeLMs learn to incorporate *compiler feedback*, *execution output*, and *traces* from self-sampled incorrect programs.
   - Demonstrated strong gains on math-reasoning code tasks over non-iterative baselines.

Professional Experience

- **Bangladesh University of Engineering and Technology (BUET)** — Dhaka, Bangladesh
  Graduate Research Assistant, Department of CSE, BUET — Mar 2021 - Present
  Supervisors: *Prof. Rifat Shahriyar and Prof. Anindya Iqbal*
  - Led projects on multilingual summarization, foundation models and benchmarks for low-resource languages and multimodal agents, culminating in **7 publications** to date.
  - Co-supervised thesis and research projects of 16 undergraduates and 2 MS students.

- **Intellesphere.AI** — Dhaka, Bangladesh
  Lead ML Engineer — Sept 2024 - Present
  - Built a Tabular QA framework for detecting anomalous transactions from financial account ledgers.
  - Developed a temporal knowledge graph-based RAG system for automatic conflict resolution of amendment chains across evolving Bengali legal documents.

- **Samsung R&D Institute** — Dhaka, Bangladesh
  Adjunct Research Assistant — Aug 2023 - Dec 2023
  - Developed a multimodal framework for automatic functional testing of Android User Interfaces.
  - Led the data curation efforts for interaction simulation datasets on internal applications.

- **Intelligent Machines Limited** Dhaka, Bangladesh
  AI Research Engineer Apr 2021 - Sep 2021
  - Created an audio annotation tool for Bengali dialects, providing on-the-fly word suggestions.
  - Built an ASR fine-tuning pipeline with Language Model integration reducing Bengali WER by 8.65%.

- **Bangladesh University of Engineering and Technology (BUET)** Dhaka, Bangladesh
  Undergraduate Research Assistant, Department of CSE, BUET Feb 2019 - Feb 2021
  Supervisor: *Prof. Rifat Shahriyar*
  - Developed *VashaBondhu*: a state-of-the-art translation system for Bengali-English, outperforming production solutions such as Google Translate and Bing Translator.
  - Curated the largest Bengali-English parallel corpora with novel sentence alignment algorithms.

Honors & Awards

- **Dean's List Award**: Bangladesh University of Engineering and Technology 2018 - 2019
- **ICT Innovation Fund for Undergraduate Thesis**: Government of Bangladesh 2020
- **University Merit Scholarship**: Bangladesh University of Engineering and Technology 2018 - 2019
- **Board Merit Scholarship**: Government of Bangladesh 2016 - 2021

Academic Services

- **Reviewer**: COLM 2025, COLING 2025, EMNLP 2024, NAACL 2024, ACL 2024, ARR 2023-25
- **Program Committee**: BNLP Workshop 2023, GEMv2 Workshop 2022

Technical Skills

- **Programming Languages**: Python, Rust, C/C++, Java, TypeScript, MATLAB, SQL, Bash, LaTeX
- **Libraries/Frameworks**: PyTorch, Keras, TensorFlow, HF Transformers, NLTK, Scikit-learn, Flask, Next.js
- **Tools/Platforms**: Docker, Git, Kubernetes, Amazon Web Services, Google Cloud Platform

Selected Courses

- Artificial Intelligence
- Machine Learning
- Computer Networks
- Probability & Statistics
- Advanced Algorithms
- Operating Systems

Reference

Rifat Shahriyar
Professor
Department of CSE, BUET
Email: rifat@cse.buet.ac.bd

Anindya Iqbal
Professor
Deparment of CSE, BUET
Email: anindya@cse.buet.ac.bd

Wasi Uddin Ahmad
Senior Research Scientist
NVIDIA AI
Email: wasiuddina@nvidia.com