Nishat Raihan

✉ mraihan2@gmu.edu   ▢ (571) 6681661   🌐 [Personal Website](#)

Summary

PhD candidate, focused on democratizing AI by creating the foundational tools to support and evaluate multilingual Large Language Models for code generation in low-resource settings. My goal is to advance the capabilities and accessibility of Code LLMs by building novel benchmarks, training corpora, and open-source models. My teaching methods are guided by my research on '*AI in CS Education*'.

EDUCATION

George Mason University   **Fairfax, Virginia**
PhD in Computer Science   *August, 2021 - May, 2026*
- Current CGPA: **3.96/4.00**

George Mason University   **Fairfax, Virginia**
MS in Computer Science   *August, 2021 - December, 2024*
- CGPA: **3.97/4.00**
- Distinguished Academic Achievement Award (CGPA>3.90)

RESEARCH

As Organizing Committee Member
- The First Workshop on Language Models for Low-Resource Languages @ **COLING'25**.
- First Workshop on Bangla Language Processing @ **EMNLP'23**.
- Multilingual Lexical Simplification Pipeline (MLSP) shared task @ **MLSP 2024**.
- Second Workshop on Bangla Language Processing @ **AACL'25**.
- 3rd International Conference on Foundation and Large Language Models @ **AACL'25**.

As Reviewer

ARR'24 & '25(Each Cycle); LREC'26; SIGCSE'24'25'26; LREC-COLING'24'25; SEMEVAL'24

As Author (Google Scholar)
- Citations: 306
- h-index: 9
- i10-index: 8

2025
- "**TigerLLM - A Family of Bangla Large Language Models**"; Nishat Raihan, Marcos Zampieri; The 63rd Annual Meeting of the Association for Computational Linguistics, (**ACL 2025**). **[PDF]**.
- "**mHumanEval - A Multilingual Benchmark to Evaluate Large Language Models for Code Generation**"; Nishat Raihan, Antonios Anastasopoulos, Marcos Zampieri; Proceedings of The Nations of the Americas Chapter of the Association for Computational Linguistics, (**NAACL 2025**). **[PDF]**.
- "**Mojobench: Language modeling and benchmarks for mojo**"; Nishat Raihan, Antonios Anastasopoulos, Marcos Zampieri; Findings of The Nations of the Americas Chapter of the Association for Computational Linguistics, (**NAACL 2025**). **[PDF]**.
- "**Large Language Models in Computer Science Education: A Systematic Literature Review.**"; Raihan, Nishat, et al.; Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1. (**SIGCSE-TS 2025**). **[PDF]**.

2024
- **Code LLMs: A Taxonomy-based Survey.**"; Raihan, Nishat, Christian Newman, and Marcos Zampieri; 2024 IEEE International Conference on Big Data (BigData). **IEEE Big Data, 2024**. **[PDF]**
- "**CSEPrompts: A Benchmark of Introductory Computer Science Prompts**"; Nishat Raihan, Dhiman Goswami, Sadiya Sayara Chowdhury Puspo, Christian Newman, Tharindu Ranasinghe,

Marcos Zampieri; 27TH INTERNATIONAL SYMPOSIUM ON METHODOLOGIES FOR INTELLI-GENT SYSTEMS (ISMIS'2024). [PDF]

o "MentalHelp: A Multi-Task Dataset for Mental Health in Social Media"; Md Nishat Raihan, Sadiya Sayara Chowdhury Puspo, Shafkat Farabi, Ana-Maria Bucur, Tharindu Ranasinghe, Marcos Zampieri; The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING'2024. [PDF]

o "Py-holmes: Causal Testing for Deep Neural Networks in Python."; McQueary, Wren, et al. , Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering. 2024. [PDF]

o "The BEA 2024 shared task on the multilingual lexical simplification pipeline."; Shardlow, Matthew, et al., Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024). 2024. [PDF]

o "An extensible massively multilingual lexical simplification pipeline dataset using the MultiLS framework."; Shardlow, Matthew, et al., Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI)@ LREC-COLING 2024. 2024. [PDF]

o "EmoMix-3L: A Code-Mixed Dataset for Bangla-English-Hindi Emotion Detection." ; Raihan, Nishat, et al., LREC-COLING 2024 (2024): 11. [PDF]

o "MasonTigers at SemEval-2024 Task 9: Solving Puzzles with an Ensemble of Chain-of-Thoughts"; Md Nishat Raihan, Dhiman Goswami, Al Nahian Bin Emran, Sadiya Sayara Chowdhury Puspo, Amrita Ganguly, Marcos Zampieri; The 18th International Workshop on Semantic Evaluation (SemEval@NAACL'2024). [PDF]

o "MasonTigers at SemEval-2024 Task 1: An Ensemble Approach for Semantic Textual Relatedness"; D Goswami, SSC Puspo, MN Raihan, ANB Emran, A Ganguly, M Zampieri; The 18th International Workshop on Semantic Evaluation (SemEval@NAACL'2024). [PDF]

o "MasonTigers at SemEval-2024 Task 8: Performance Analysis of Transformer-based Models on Machine-Generated Text Detection"; SSC Puspo, MN Raihan, D Goswami, ANB Emran, A Ganguly, O Uzuner; The 18th International Workshop on Semantic Evaluation (SemEval@NAACL'2024). [PDF]

o "MasonPerplexity at ClimateActivism 2024: Integrating Advanced Ensemble Techniques and Data Augmentation for Climate Activism Stance and Hate Event Identification"; A Ganguly, SSC Puspo, D Goswami, MN Raihan; Fourth Workshop on Language Technology for Equality, Diversity, Inclusion (LTEDI@EACL'2024). [PDF]

o "MasonPerplexity at Multimodal Hate Speech Event Detection 2024: Hate Speech and Target Detection Using Transformer Ensembles"; A Ganguly, ANB Emran, SSC Puspo, MN Raihan, D Goswami, M Zampieri; 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE@EACL'2024). [PDF]

o "MasonTigers@ LT-EDI-2024: An Ensemble Approach towards Detecting Homophobia and Transphobia in Social Media Comments"; D Goswami, SSC Puspo, MN Raihan, ANB Emran, 2024. LT-EDI 2024 : Fourth Workshop on Language Technology for Equality, Diversity, Inclusion (LTEDI@EACL'2024). [PDF]

2023

o [BEST PAPER AWARD] "Offensive Language Identification in Transliterated and Code-Mixed Bangla"; Md Nishat Raihan, Umma Hani Tanmoy, Anika Binte Islam, Kai North, Tharindu Ranasinghe, Antonios Anastasopoulos, Marcos Zampieri. Proceedings of the 1st International Workshop on Bangla Language Processing (BLP-2023), EMNLP, Singapore. [PDF] [DATASET]

o "nlpBDpatriots at BLP-2023 Task 1: A Two-Step Classification for Violence Inciting Text Detection in Bangla"; Md Nishat Raihan, Dhiman Goswami, Sadiya Sayara Chowdhury Puspo, Marcos Zampieri. Proceedings of the 1st International Workshop on Bangla Language Processing (BLP-2023), EMNLP, Singapore. [PDF]

o "nlpBDpatriots at BLP-2023 Task 2: A Transfer Learning Approach to Bangla Sentiment Analysis"; Dhiman Goswami, Md Nishat Raihan, Sadiya Sayara Chowdhury Puspo, Marcos Zampieri. Proceedings of the 1st International Workshop on Bangla Language Processing (BLP-2023), EMNLP, Singapore. [PDF]

o "OffMix-3L: A Novel Code-Mixed Dataset in Bangla-English-Hindi for Offensive Language Identification"; Dhiman Goswami, Md Nishat Raihan, Antara Mahmud, Antonios Anastasopoulos,

Marcos Zampieri. The First Workshop in South East Asian Language Processing, AACL 2023, Bali, Indonesia. **[PDF] [DATASET]**

o "SentMix-3L: A Bangla-English-Hindi Code-Mixed Dataset for Sentiment Analysis"; Md Nishat Raihan, Dhiman Goswami, Antara Mahmud, Antonios Anastasopoulos, Marcos Zampieri. The 11th International Workshop on Natural Language Processing for Social Media. AACL 2023. Bali, Indonesia. **[PDF] [DATASET]**

o "Mixed-Distil-BERT: Code-mixed Language Modeling for Bangla, English, and Hindi."; Raihan, Md Nishat, Dhiman Goswami, and Antara Mahmud. arXiv preprint arXiv:2309.10272 (2023). **[PDF] [MODEL]**

o "Determining the Optimal Number of Clusters for Time Series Datasets with Symbolic Pattern Forest"; Raihan, Md Nishat. arXiv preprint arXiv:2310.00820 (2023). **[PDF]**

2022..

o "An Experimental Analysis on the Sensitivity of the Most Widely Used Edge Detection Methods to Different Noise Types"; Raihan, M., Ulfat, N, and Saqib, Nazmus, March 10-12, 2022. International Conference on Computing Advancements (ICCA'22). **[PDF]**

2020..

o "A Novel Approach to Classify Electrocardiogram Signals Using Deep Neural Networks", Ahmed, T., Rahman, A., Chowdhury, T.M., Kushol, R. and Raihan, M.N., 2020, October. In 2020 2nd International Conference on Computer and Information Sciences (ICCIS). IEEE. **[PDF]**

2019..

o "Contrast Enhancement of Medical X-Ray Image Using Morphological Operators with Optimal Structuring Element"Kushol, R., Raihan, M., Salekin, M.S. and Rahman, A.B.M., 2019. arXiv preprint arXiv:1905.08545. **[PDF]**

o "A Complete Bangla Optical Character Recognition System: An Effective Approach" T. Ahmed, M. N. Raihan, R. Kushol and M. S. Salekin, 2019 22nd International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2019, pp. 1-7. **[PDF]**

2018..

o "An Effective Navigation System Combining both Object Detection and Obstacle Detection Based on Depth Information for the Visually Impaired" Raihan, Md, and Hossain Mohammad Seym. Diss. Department of Computer Science and Engineering, 2018. **[PDF]**

o "An Android-Based Useful Text Extraction Framework Using Image and Natural Language Processing", Rafsanjany Kushol, Imamul Ahsan, and Md. Nishat Raihan, International Journal of Computer Theory and Engineering (IJCTE), vol. 10, no. 3, pp. 77-83, 2018. **[PDF]**

Datasets..

1. **mHumanEval-Benchmark:** The mHumanEval benchmark includes a total of 836,400 coding prompts and covers 204 different natural languages and 25 programming languages. **[Github]**
2. **Bangla-Instruct:** A 400K High-Quality Instruction dataset, for finetuning Bangla Large Language Models. **[Github] [HuggingFace]**
3. **Bangla-TextBook:** A 10M token corpus, generated from Bangla Textbooks for pretraining Bangla Large Language Models. **[Github] [HuggingFace]**
4. **HumanEval-Mojo:** A Benchmark to evaluate Large Language Model on Mojo code generation. **[Github]**
5. **Mojo-SFT:** An instruction dataset for finetuning Large Language Models to generate Mojo Code. **[Github]**
6. **Mojo-*mSFT*:** A *multilingual* instruction dataset for finetuning Large Language Models to generate Mojo Code. **[Github]**
7. **Mojo-Corpus:** A 6.5M token corpus for pretraining Large Language Models on Mojo language. **[Github]**
8. **CSEPrompts:** A Benchmark of Introductory Computer Science Prompts; 200+ coding prompts and 100+ MCQ prompts. **[Github]**

9. **MentalHelp:** 14 million, semi-supervised, mental disorder detection data. **[Github]**
10. **TB-OLID:** Offensive Language Identification in Transliterated and Code-Mixed Bangla. **[Github]**
11. **SentMix-3L:** 1K Human-Generated Data for Sentiment Analysis, Code-Mixed (Bangla-English-Hindi). **[Kaggle] [Github]**
12. **OffMix-3L:** 1K Human-Generated Data for Offensive Language Identification, Code-Mixed (Bangla-English-Hindi). **[Kaggle] [Github]**
13. **Code Mixed Sentiment [Bangla-English-Hindi]:** 100k Sentiment Analysis Data, 3 labels, 3 languages, Code-mixed, Synthetic. **[Kaggle] [Github]**
14. **Code Mixed Offensive [Bangla-English-Hindi]** 100k Offensive Language Identification Data, 2 labels, 3 languages, Code-mixed, Synthetic. **[Kaggle] [Github]**

Models

1. **TigerCoder:** The SOTA LLM-family for Bangla. **[HuggingFace]**
2. **Mojo-Coder Family:**
   o *Mojo-Coder:* Base Large Language Model for Mojo.
   o *Mojo-Coder-it:* Finetuned Large Language Model for Mojo.
   o *Mojo-Coder-it-m:* Multilingually finetuned Large Language Model for Mojo.
3. **Tri-Distil-BERT:** Pretrained on Bangla-English-Hindi. **[HuggingFace]**
4. **Mixed-Distil-BERT:** Pretrained Tri-Distil-BERT on 560k Code-mixed data. **[HuggingFace]**

PROFESSIONAL EXPERIENCES

o **George Mason University** **Fairfax, Virginia**
Graduate Research Assistant, Computer Science *May, 2024 - Present*

Responsibilities:

- Conduct supervised research, including literature reviews, data collection, and analysis, and report writing.
- Assist in coordinating and managing research projects to meet deadlines.
- Responsible for the collection, management, and meticulous analysis of data sets relevant to the research.
- Presentations: Compile, organize, and present research findings at internal meetings, academic conferences, or other public forums.
- Collaboration: Collaborate efficiently with other research team members, contributing to a dynamic and productive research environment.
- Publication: Participate in the drafting, editing, and submission of academic papers and articles based on the research findings.

o **George Mason University** **Fairfax, Virginia**
Graduate Teaching Assistant *August, 2021 - May, 2024*
Courses: CS 222 (Computer Programming for Engineers),
CS 262 (Intro. to Low-Level Programming)
& CS 583 (Analysis of Algorithms)

Responsibilities:

- Conducting lab sessions and lab recitations.
- Grading assignments on a regular basis.
- Helping the students with their mistakes.
- Helping the students with after-class issues.
- Assisting the Professor to monitor exams.

o **Samsung R&D Institute Bangladesh** **Dhaka, Bangladesh**
Software Engineer, RPA & OCR Development *November, 2019 - August, 2021*

Responsibilities:

- Worked with RPA tools - Automation Anywhere, UI Path.
- Developed complex RPA services and bots for multiple international organizations.
- Worked with existing OCR engines - Tesseract4, Google Vision API, Abbyy, Microsoft Azure API

and Tegaki API.
- Developed IQ Bots for OCR purposes.
- Wrote Python scripts and integrated them with the IQ bots to work efficiently under most circumstances.
- Worked with OpenText and SAP.

Platforms: Automation Anywhere, UI Path, OpenText, SAP.

o **Uttara University**   **Dhaka, Bangladesh**
Lecturer, Computer Science   *January, 2019 - April, 2019*

Responsibilities:
- Taught theory courses - Introduction to Computers, C, C++.
- Conducted Lab courses - C, C++.
- Participated in research works - Computer Vision.
- Mentored students for their betterment.

VOLUNTARY ACTIVITIES

o **Member**   **George Mason Chapter**
The National Society of Leadership and Success (NSLS)   *Oct, 2024 - Present*

o **Vice-President**   **Fairfax, Virginia**
Bangladesh Graduate Student Association (BDGSA)   *Apr, 2023 - Mar, 2024*

o **Director of Sports**   **Fairfax, Virginia**
Bangladesh Graduate Student Association (BDGSA)   *Apr, 2022 - Mar, 2023*

REFERENCES

Attached Separately

****