M SAIFUL BARI

sbmaruf.github.io
sbmaruf@gmail.com

EDUCATION

Nanyang Technological University *Jan 2019 - Aug 2023*
Doctor of Philosophy (Ph.D.)
Natural Language Processing
Computer Science and Engineering

EXPERIENCE

Senior Research Scientist Aug 2023 - Present
National Center for Artificial Intelligence

- **I'm one of the Core Maintainer and Research Lead of ALLaM-** a nationwide LLM effort of Saudi Arabia.
- Currently working on the Pre-training, Leading Instruction Tuning and the Alignment of the Foundation Models.
- I lead the alignment team at SDAIA. My lab's current objective is to improve the cross-lingual linguistic alignment.
- I'm investigating alignment at scale during pretraining to reduce the distributional shift and alignment discrepancy between objectives.

Applied Scientist Intern July 2022 - Nov 2022
Amazon Development Service

- Worked on parameter efficient multitask inference (PEMI) training in Large Language Model.

BLOOM LLM training Sep 2021 - Mar 2022
International Research Effort

- Worked on large-scale LLM training in Architecture WG and Prompt Engineering WG.

Applied Scientist Intern (Part Time) July 2021 - Jan 2022
Amazon Web Service

- Working on prompt tuning for multi-lingual models.

Applied Scientist Intern Aug 2020 - Oct 2020
Amazon Web Service

- Worked on Cross-lingual Few-shot Adaptation.

Research Assistant Sep 2017 - Aug 2018
Nanyang Technological University

- Research on MT, NER and Adversarial Training.

Aubichol Intelligent Technologies Sep 2017 - Aug 2018
Product Development Intern

- Help an early *Sports Analytic* start-up to build their MVP.

Software Engineering Intern Nov 2015 - Dec 2015
XeonBD

- The course of the internship goes through Cloud Computing and kernel virtualization.

HONORS & AWARDS

Scholarship 2019
NTU Research Scholarship, Fully funded Ph.D. scholarship for 4 years.

Scholarship 2012
OIC Scholarship for undergraduate study, Islamic University of Technology

Champion 2014
IUT Computer Programming Contest

Honorable Mention 2014
Human Expedition on Mars Timeline 2018

Champion 2015
IUT Computer Programming Contest

2nd/100 in Inter University Programming Contest 2016
Daffodill International University ACM ICPC world finals warmup contest 2016

6th/100+ in Inter University Programming Contest 2016
NSU Cybernauts National Programming Contest

SKILLS

Computer Languages	Python, C/C++, Bash
Tools & Library	Pytorch, Megatron-LM, NeMo, Huggingface, DeepSpeed, vllm
Problem Solving	Deep Learning, Algorithms & Data Structure

PUBLICATION

Summary: ACL-19, AAAI-20, EMNLP-20, 2*ACL-21, EMNLP-21, ICLR-22, ACL-22, EMNLP-22, 3*ACL-23, 3 Preprint, 1 Book Chapter

1. **M Saiful Bari***, Yazeed Alnumay*, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan AlRashed, Faisal A. Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, Ahmed Abdelali, Jeril Kuriakose, Abdalghani Abujabal*, Nora Al-Twairesh*, Areeb Alowisheq*, and Haidar Khan*. **ALLaM: Large Language Models for Arabic and English**, 2024

2. Md Tahmid Rahman Laskar, Sawsan Alqahtani, **M Saiful Bari**, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, Enamul Hoque, Shafiq Joty, and Jimmy Huang. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations, 2024

3. **M Saiful Bari***, Mohammad Abdullah Matin* Khan, Xuan Long Do, Weishi Wang, Md Rizwan Parvez, and Shafiq Joty. xCodeEval: **A Large Scale Multilingual Multitask Benchmark for Code Understanding, Generation, Translation and Retrieval**, 2024

4. Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, Nouf Alotaibi, Nora Altwairesh, Areeb Alowisheq, **M Saiful Bari**, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards, 2024

5. **Bari, M Saiful***, Laskar Tahmid*, Rahman Mizanur, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. **A Systematic Study of ChatGPT on Benchmark Datasets**. In *Findings of the 61th Annual Meeting of the Association for Computational Linguistics, ACL*, 2023

6. **M Saiful Bari**, Aston Zhang, Shuai Zheng, Xingjian Shi, Yi Zhu, Shafiq Joty, and Mu Li. **SPT: Semi-Parametric Prompt Tuning for Multitask Prompted Learning** (under review at the 2023 conference on empirical methods in natural language processing, emnlp'23), 2022

7. Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, **M Saiful Bari**, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. **Crosslingual Generalization through Multitask Finetuning**. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics, ACL*, 2023

8. Zheng-Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, **M Saiful Bari**, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Indra Winata, Stella Biderman, Dragomir Radev, and Vassilina Nikoulina. **BLOOM+1: Adding Language Support to BLOOM for Zero-Shot Prompting**. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics, ACL*, 2023

9. Teven Le Scao, Thomas Wang, Daniel Hesslow, Lucile Saulnier, Stas Bekman, **M Saiful Bari**, Stella Biderman, Hady Elsahar, Jason Phang, Ofir Press, Colin Raffel, Victor Sanh, Sheng Shen, Lintang Sutawika, Jaesung Tae, Zheng Xin Yong, Julien Launay, and Iz Beltagy. **What Language Model to Train if You Have One Million GPU Hours?** In *Findings in EMNLP, 2022*, 2022

10. Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, **M Saiful Bari**, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan Fries, Maged S. Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-Jian Jiang, and Alexander M. Rush. **PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts**. In *Meeting of the Association for Computational Linguistics (ACL) Demonstration*, 2022

11. Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, **M Saiful Bari**, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. **Multitask Prompted Training Enables Zero-Shot Task Generalization**. In *International Conference on Learning Representations, ICLR*, 2022

12. Teven et al, **BLOOM: A 176B-Parameter Open-Access Multilingual Language Model**, ArxiV preprint, (Under Review at Journal of Machine Learning Research, JMLR), 2022

13. **Bari, M Saiful**, Batool Haider, and Saab Mansour. **Nearest Neighbour Few-Shot Learning for Cross-lingual Classification**. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1745–1753, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics

14. Tasnim Mohiuddin, **M Saiful Bari**, and Shafiq Joty. **AugVic: Exploiting BiText Vicinity for Low-Resource NMT**. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online, 2021. Association for Computational Linguistics

15. **Bari, M Saiful**, Tasnim Mohiuddin, and Shafiq Joty. **UXLA: A Robust Unsupervised Data Augmentation Framework for Zero-Resource Cross-Lingual NLP**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL (Volume 1: Long Papers)*, pages 1978–1992, Online, August 2021. Association for Computational Linguistics

16. Tasnim Mohiuddin, **M. Saiful Bari**, and Shafiq R. Joty. **LNMap: Departures from Isomorphic Assumption in Bilingual Lexicon Induction Through Non-Linear Mapping in Latent Space**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Virtual, November 2020

17. **M Saiful Bari**, Shafiq Joty, and Prathyusha Jwalapuram. **Zero-Resource Cross-Lingual Named Entity Recognition**. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, AAAI '20, New York, USA, 2020. AAAI

18. Xiang Lin, Shafiq Joty, Prathyusha Jwalapuram, and **M Saiful Bari**. **A Unified Linear-Time Framework for Sentence-Level Discourse Parsing**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL '19, Florence, Italy, 2019. ACL

19. **M Saiful Bari**. ***Regression for Data Analytics***, chapter 2, pages 33–54. CRC Press, Boca Raton, September 2018. (in book: Data Analytics: Concepts, Techniques and Applications)

PROJECTS

⋆ ExecEval :: *Evaluation Framework*
Details can be found here, https://github.com/ntunlp/ExecEval

⋆ Data Sampling :: *xCodeEval*
Details can be found here, https://github.com/ntunlp/xCodeEval

⋆ Prompt Engineering :: *Promptsource*
Details can be found here, https://github.com/bigscience-workshop/promptsource

⋆ LLM Pipelining :: *Megatron/DeepSpeed*
Details can be found here, https://github.com/bigscience-workshop/Megatron-DeepSpeed

⋆ Dataset :: *xCodeEval*
Details can be found here, https://github.com/ntunlp/xCodeEval

⋆ Deep learning :: *UXLA*
Details can be found here, https://github.com/sbmaruf/UXLA

⋆ Deep learning :: *Cross-lingual Few Shot Learning*
Details can be found here, https://github.com/amazon-science/nearest-neighbor-crosslingual-classification

⋆ Deep learning :: *Zero-Resource Cross-lingual Named Entity Recognition*
Details of the project can be found here, https://github.com/ntunlp/Zero-Shot-Cross-Lingual-NER

⋆ Deep learning :: *Malay English Machine Translation System*
Details can be found here, https://sbmaruf.github.io/project/mt-system/

⋆ Machine learning system :: *A CBIR System*
Details of the project can be found here, http://103.82.172.44:8080/xmlui/handle/123456789/93

⋆ Algorithms :: *Algorithm-Code-Library*
Details of the project can be found here, https://github.com/sbmaruf/Algorithms-Code-Library