Md Mushfiqur Rahman

571.668.1664 | mrahma45@gmu.edu | mushfiqur11 | mushfiqur11
https://mushfiqur11.github.io | Google Scholar Profile

PROFILE

- 4th year PhD student with a Master's degree in Computer Science (Machine Learning concentration) and a strong background in Mathematics. Specializes in NLP and its applications in health. Experienced in building and managing industry-standard ML pipelines.
- Published 8 peer-reviewed papers, including 2 in Q1 Journals and 2 at ACL workshops.
- Collaborated with the US Naval Research Laboratory to build a GAN-based image denoiser that is actively being used in the Hinode spacecraft.

CORE EXPERTISE

ML Pipelines	**Generative AI**	**Backend & Database**	**NLP in Health**
AI Chatbots	**Cloud**	**User Study**	**Continual Learning**
Cross-lingual Transfer	**Image Denoising**	**Game Development**	**Text Simplification**

TECHNICAL SKILLS AND CERTIFICATIONS

Languages: Python, Java, C#, C++, Octave, TypeScript (Angular), JavaScript (React)
ML Tools: PyTorch, TensorFlow, Scikit-learn, GANs, RAG, Reinforcement Learning (RL and RLHF), Code2Vec
NLP & Vision: Hugging Face, LangChain, On-device LLM, Fine-tuning, GPT, OpenCV | **Visualization:** Matplotlib, Seaborn
Cloud: AWS (Lambda, EC2, S3), Google Cloud Platform (GCP), Firebase, Azure | **Platforms:** Web, Windows, Linux
Backend: REST, JUnit, FastAPI, Docker | **Database:** SQL, MongoDB, MySQL, NoSQL, SQLite, PL/SQL
Tools: Cloud console, GitHub Enterprise, CI/CD, Postman, AI Evaluation, HCI, Developer Productivity Tools
Certifications: "Deep Learning Specialization" (5 courses by Deeplearning.AI),
"Machine Learning" (by Stanford University through Coursera),
"Front-end JS Framework: Angular" (by HKUST through Coursera)

EXPERIENCE

George Mason University — **Fairfax, VA**
Graduate Research Assistant — Aug. 2021 – Present

- Research on NLP with a focus on improving information accessibility in the healthcare domain. Published **5 research papers**.
- Investigating how tone and dialect variations impact the performance and robustness of LLMs, with the goal of developing models that can effectively address these challenges
- Studying LLM performance in key tasks such as conversational agents, information retrieval, and question-answering systems

US Naval Research Laboratory — **SW Washington, DC**
Machine Learning Software Developer (Intern) — May 2024 – Aug. 2024

- Designed a **Retrieval-Augmented Generation** pipeline, processing 100K+ technical documents to achieve real-time response rates.
- Engineered the embedding and retrieval framework using **MiniLMEmbedder** and **WindowRetriever**, reducing the memory footprint and computational load by 20%.
- Integrated Llama-3.1-70B as the generative model for a seamless interaction with the RAG system.

Machine Learning Software Developer (Intern) — May 2023 – Aug. 2023

- Built a **GAN (Generative Adversarial Network)-based image denoiser** for the EIS (Extreme-ultraviolet Image Spectrometer) machine hosted in the Hinode spacecraft.
- The ML model is **actively being used in the Hinode spacecraft** and **removes 93% noise** from the EIS-generated images. It achieves **0.48 average test error** which is significantly better than the previous best (0.72).

Military Institute of Science and Technology — **Dhaka, Bangladesh**
Faculty (Lecturer) — Mar 2021 – Aug. 2021

- Courses: "Introduction to Artificial Intelligence" and "Algorithms I"
- Co-supervised **3 undergraduate thesis projects** related to machine learning.

Samsung R&D Institute Bangladesh — **Dhaka, Bangladesh**
ML Research Intern — Nov. 2019 Jan. 2020

- Developed a deep learning model that generates **3D objects from equirectangular (box-map) images**.
- The GAN-like model (adjusted for spherical images) achieved **0.21 RMS error** for the 3-D vertices

M-World — **Dhaka, Bangladesh**
Game Development (Team Lead) — Feb. 2019 Sept. 2019

- Led a cross-functional team of 2 developers, a designer, and a QA tester to develop a **hyper-casual mobile game** (for Android).
- Managed the complete **software development lifecycle**, from concept and design to deployment and post-launch support.
- Utilized **Agile methodologies**, leading daily stand-ups and sprint planning meetings to ensure project milestones were met on schedule.

EDUCATION

George Mason University — **Fairfax, VA**
PhD in Computer Science (CGPA: 3.90) — Aug. 2021 – May 2026

- Master's (MSc) in Computer Science (with Machine Learning concentration) (Degree received on May 2024)

Islamic University of Technology — **Gazipur, Bangladesh**
Bachelor's (BSc) in Computer Science and Engineering (CGPA: 3.79) — Jan. 2017 – Feb. 2021

PROJECTS

Spectral Reconstruction of EIS Images with Conditional GAN

- Compared 4 image reconstruction/denoising algorithms (GAN, cGAN, WGAN, and Autoencoder) for EIS images.
- The cGAN **removes 93% noise** and improves image denoising capabilities of the EIS hosted in Hinode.

Health Information Simplification and Summarization [Code] [Paper]

- Created a parallel text dataset for **health-text simplification** in collaboration with nurse practitioners and patient-communication experts.
- Fine-tuned **GPT4** and **Llama2** (with **PPO** and **Reinforcement Learning**) on this dataset to get SOTA results in the domain.
- Trained Llama-2 with a 16-bit quantized Q-LoRA (minimizing computational resource requirement by **60% for GPU usage and 90% for time**). Reduced data requirements significantly by introducing a novel **Supervised Fine-tuning (SFT)**+RLHF model (with a novel reward).

Modes in Cross-lingual Transfer (tokenization vs non-tokenization) [Code] [Paper]

- Compared tokenization-based models (`BERT`, `mBERT`) with tokenization-free models (`PIXEL`, `CANINE`). Proposed a decision aid for text-representation based on **19 source and 133 target languages** for **3 common NLP tasks**.
- The analysis underscores the **importance of non-tokenization models** for specific tasks.

Mitigating Catastrophic Forgetting in Continual Learning

- Developed an algorithm to mitigate catastrophic forgetting enabling ML models to retain high performance on previously learned tasks.
- The algorithm uses **A-GEM** as its base but equates all higher-degree polynomials in the second derivative of the Taylor series as zero.

COVID Misinformation Detection

- Built a BERT-based model for identifying misinformation across diverse datasets. **Accuracy:** 97% on Constraint2021 and 93% on CoAID
- Extensively analyzed the effectiveness of **LLMs (GPT3) in zero-shot and few-shot settings** for this task.

Intent Detection and Slot Filling for Home Assistants [Code] [Paper]

- Built the **first ever intent detection and slot filling dataset** for Bangla and Sylheti.
- GPT-3.5 model achieves **0.94 (intent detection) and 0.51 (slot filling)** F1 score for Bangla.

Analyzing the Effect of Attentions in Video Captioning [Code] [Paper]

- Compared different attention mechanisms and for video captioning. Created a **video captioning model** with stacked attention that achieved **state-of-the-art performance on MSVD**.
- A **detailed ablation study of each component** of the model showed the contribution of the stacked attention in retaining semantic information.

Image In-painting and Reconstruction with GAN [Code]

- Proposed a **novel loss function** for image in-painting that utilizes the structural information and edge-map from images.
- Our method, **StructGAN**, enhances image restoration by using the structure loss, achieving 16% higher SSIM scores than SOTA.

AWARDS AND SCHOLARSHIPS

- **Distinguished Academic Achievement Award**, from George Mason University (2024)
- **Runner-up of Inter-University App Development Contest**, IUT ICT Fest (2019)
- **Champion of ICT4D** at 4th AUW ICT Fest (2018)
- **National Topper** in Mathematics in IAS (International Assessment for Schools) by UNSW (x2) (2008, 2014)
- **National Mathematical Olympiads** Medals (x4) (2009, 2012, 2013, 2014)

SELECTED PUBLICATIONS

[1] [Q1 Journal] **Md Mushfiqur Rahman**†, Mohammad Sabik Irbaz†, Kai North, Michelle S. Williams, Marcos Zampieri, and Kevin Lybarger. 2024. Health text simplification: An annotated corpus for digestive cancer education and novel strategies for reinforcement learning. *Journal of Biomedical Informatics*, 158:104727

[2] **Md Mushfiqur Rahman**, Fardin Ahsan Sakib, Fahim Faisal, and Antonios Anastasopoulos. 2023. To token or not to token: A comparative study of text representations for cross-lingual transfer. In *Workshop on Multi-lingual Representation Learning (MRL)*. Association for Computational Linguistics

[3] Fardin Ahsan Sakib†, A H M Rezaul Karim†, Saadat Hasan Khan†, and **Md Mushfiqur Rahman**†. 2023. Intent detection and slot filling for home assistants: Dataset and analysis for Bangla and Sylheti. In *Workshop on Bangla Language Processing (BLP)*. Association for Computational Linguistics

[4] Bin Han†, Haotian Zhu†, Sitong Zhou†, Sofia Ahmed, **Md Mushfiqur Rahman**, Fei Xia, and Kevin Lybarger. 2023. Huskyscribe at mediqa-sum 2023: Summarizing clinical dialogues with transformers. Thessaloniki, Greece. CLEF

[5] Refaat Mohammad Alamgir, Ali Abir Shuvro, Mueeze Al Mushabbir, Mohammed Ashfaq Raiyan, Nusrat Jahan Rani, **Md Mushfiqur Rahman**, Md. Hasanul Kabir, and Sabbir Ahmed. 2022. Performance analysis of yolo-based architectures for vehicle detection from traffic images in bangladesh. In *25th International Conference on Computer and Information Technology (ICCIT)*

[6] [Q1 Journal] **Md Mushfiqur Rahman**†, Thasin Abedin†, Khondokar SS Prottoy, Ayana Moshruba, and Fazlul Hasan Siddiqui. 2021. Video captioning with stacked attention and semantic hard pull. *PeerJ Computer Science*

[7] **Md Mushfiqur Rahman**, Sabah Binte Noor, and Fazlul Hasan Siddiqui. 2020. Automated large-scale class scheduling in minizinc. In *International Conference on Sustainable Technologies for Industry 4.0 (STI)*

INVITED TALKS

- May 2024, Naval Research Laboratory: “Leveraging Generative AI for Advanced Scientific Research: Breakthroughs and Barriers”

Equally contributing authors marked with †.