

Nazia Tasnim

✉ nimzia@bu.edu
🌐 appledora.github.io [in LinkedIn](#) [Google Scholar](#) [appledora](#)
🏠 Boston, MA.

EDUCATION

Boston University, MA.

PhD. Student in Computer Science (September'23 - Present)

Advisor: Dr. Bryan Plummer

RESEARCH INTEREST

Primary: Parameter-efficient Finetuning, Model Editing, Adversarial Attacks, Machine Unlearning
Emphasis: Bias Mitigation & Fair Representations, Robust Domain Adaptation, Knowledge Infusion, AI For Social Good

PUBLICATIONS

1. RECAST : Reparameterized, Compact weight Adaptation for Sequential Tasks | *Under Review*
 - Proposed a novel method that reduces the number of task-specific trainable parameters to fewer than **50**. We empirically demonstrated that RECAST outperforms the state-of-the-art by up to **3%** across various scales, architectures, and parameter spaces
2. [Vision-LLMs Can Fool Themselves with Self-Generated Typographic Attacks](#) | *NeuRIPS'24*
 - Introduced **two** novel self-generated attacks that prompt the LVLM to generate an attack against itself. Our benchmark reveals that Self-Generated attacks can reduce LVLM's classification performance by up to **33%**
3. [OOD-Speech: A Large Bengali Speech Recognition Dataset for Out-of-Distribution Benchmarking](#) | *INTERSPEECH'23*
 - Collected **1177.94** hours collected and curated from 22, 645 native Bengali speakers, with additional **23.03** hours of speech collected and manually annotated from 17 different sources. Jointly the largest and only OOD dataset for Bengali.
4. [VISTA: Vision transformer enhanced by U-Net and image colorfulness frame filtration for automatic retail checkout](#) | *IEEE/CVF Conference on CVPR'22*
 - Implemented an end-to-end pipeline achieving **46%** accuracy in real-time multi-class product recognition, the *3rd* highest score for the cross-modality dataset.
5. [On leveraging data augmentation and ensemble to recognize complex Named Entities in Bangla](#) | *The International Workshop on SemEval'21*
 - Implemented 3 different model ensembles and generated 6 augmented datasets splits. The final approach obtained **60%** f1-score, the *8th* highest performance in the dataset.
6. [Exploring the Scope and Potential of Local Newspaper-based Dengue Surveillance in Bangladesh](#) | *KDD Workshop on Applied Data Science for Healthcare'21*.
7. [Observing the Unobserved: A Newspaper-Based Dengue Surveillance System for the Low-Income Regions of Bangladesh](#) | *The International FLAIRS Conference Proceedings'21*.
8. [Choice of assemblers has a critical impact on de novo assembly of SARS-CoV-2 genome and characterizing variants](#) | *Briefings in Bioinformatics*.

CURRENT RESEARCH

IVC-ML Group, Boston University

Sept'23 - Present

Graduate Researcher

- Working on model decomposition and reconstruction to support resource-bounded settings.
- Developed efficient **Reparameterization Schemes** for incremental learning that increases image classification performance by up to 3% with $< 2 * 10^{-6}$ tunable parameters.
- Developed novel **probes and attacking techniques** to analyze vulnerabilities in *LVLMS*

EXPERIENCES

Wikimedia Foundation

Sept'22 - June'23

Research Developer

- Collaborated with Wikimedia's research team to **develop NLP tools** for 300+ languages in 100+ wikiprojects
- Created research pipeline to **curate community resources** from wiki projects, developed assets through analysis, build models and off-the-shelf tools to be used across all wiki research

Giga Tech Limited

Aug'21 - June'22

Machine Learning Engineer

- Developed NLP submodules for the **National Syntactic Treebank**
- Created pipelines for **Part-of-Speech (PoS), Named Entity Recognition (NER)** and **language model training**
- Assisted in establishing annotation guidelines for diverse downstream tasks

Bengali.AI

Sept'21 - Present

Research Affiliate

- **Coordinated teams, lead research projects** and published multiple research papers focused on alleviating the low-resource status of Bengali
- Launched **Google-funded Kaggle competitions** and inter-university DL contests with 600+ participants
- Developed some of the **largest benchmarking datasets for Bangla NLP**

Newsroom Lab, BRAC University

Sept'22 - June'23

Research Assistant

- Led a team of undergraduate students in building tools for **social science data analysis**
- Developed an end-to-end pipeline to **identify primary speakers in news clusters** and establish their geopolitical affiliations
- Curated a dataset to generate **ontology of geopolitical association** by combining metainformation from Wikipedia and Wikidata.

KATclub.ca

Coding Instructor

Oct'20 - July'21

Instructed middle to high school students to start with problem-solving. Taught **web development, game development, and interactive problem solving**

SELECTED PROJECTS

mwtokenizer [[Package](#)]

A multilingual **Python tokenization package** for Wikimedia Projects focusing on non-whitespace delimited languages. Analyzed large-scale wikicorpora through **PySpark** to optimize pattern recognition in diverse writing systems and curated essential module assets.

mwparserfromhtml [[Package](#)]

A **Python** library to parse and extract metadata from Enterprise HTML Dumps. The module is part of the core components of the **Mediawiki Utilities Project**.

VISTA [[Code](#)]

Developed an end-to-end pipeline for real-time retail checkout, combining **UNET**-based segmentation with *entropy masking* for domain bias reduction. Implemented multi-class product classification using **Vision Transformers (ViT)**. Designed a **custom metric** for frame selection, optimizing object detection efficiency in video streams.

Shaako [[Code](#)] [[Demo](#)]

A mobile application to provide rural people with Emergency Medical Care support through community health workers. Built the frontend in **Android Framework**. The backend portal is **Django** and uses both **Firebase** and **Azure CMS** for database.

SUSTCast [[Code](#)] [[App](#)]

Created and launched the pioneering SUST campus online radio platform, featuring fully automated streaming via an **IceCast server** accommodating concurrent engagement of *2000+ listeners*. Designed the app interface using **Firebase** real-time database, while collaborating on the development of *data collection, processing, and music recommendation*.

HONORS AND AWARDS

Recipient of Rafik B. Hariri Research Fellowship, Fall'24

Recipient of Dean's Fellowship, Boston University GSAS, Fall'2023 - Summer'2024

Outreachy Internship at the Wikimedia Foundation (top 1.4% applicant) [[Mentee](#)]

Second Runner-up at MSFT Imagine Cup - SEANM, 2022 [[Demo](#)]

Second Runner-up, CVPR AI City Challenge Track-4, 2022 (3rd globally) [[Code](#)]

Grace Hopper Scholarship Recipient, 2021 (Top 5% scholars)

4th place in AI-Based Dhaka Traffic Detection Challenge, 2020 [[Poster](#)]

Champion at BRACATHON 3.0 - Healthcare Category, 2019 [[Project Pilot](#)]

LEADERSHIP AND COMMUNITY SERVICE

- Judge | **BDOSEN - NLP Hackathon, 2023**
- Organizer | **BDOSEN - Ada Lovelace Celebration, 2022**
- Founding Member & Membership Chair | **SUST ACM Students' Chapter, 2019-2020**
- Organizing Committee | **IEEE International Conference on Bangla Speech and Language Processing (ICBSLP), 2019**

TECHNICAL SKILLS

- **Languages** Python, Java, JavaScript, SQL, SPARQL
- **Framework** Android, Django, React, Flask
- **Machine Learning** PyTorch, Keras, OpenCV, pySpark
- **Research Methodology** Large-scale dataset curation, Controlled experiments and ablation, Literature review and Meta-analysis, Interdisciplinary research collaboration and communication