Harshita **Diddee**

PhD Student, Carnegie Mellon University

 Portfolio   Github   Google Scholar   @ Email

Education

August 2023 Spring 2028	**Carnegie Mellon University** PhD, Language Technologies Institute: Advised by Daphne Ippolito Working on Data Quality Estimation and Watermarking LLM Generations.	**Pittsburgh, USA**
May 2017 Jun 2021	**Guru Gobind Singh Indraprastha University** B.Tech., Computer Science & Engineering	Department Rank: 2/120 Graduated as the Best Outgoing Student for the Class of 2021	**Delhi, India**

Select Experience

Jul 2021 July 2023	**Microsoft Research** *SCAI Centre Fellow	Advisor: Dr. Kalika Bali* Developing edge-friendly machine translation models for extremely low-resource languages. Evaluating GPT across its (a) multi-lingual abilities (b) task-coverage and (c) capability as an evaluator.	**Bangalore, India**
Jun 2022 Aug 2022	**Johns Hopkins University** *Visting Pre-Doctoral Research	Host: JSALT'22* Evaluated the generalizability of speech and text cross-lingual models to extremely low-resource languages as a part of the Speech Translation for Under-Resourced Languages Track.	**Baltimore, USA**
May 2019 Oct 2019	**Indian Institute Of Technology, Delhi (IIT-D)** *Research Intern	Advisor: Aakanksha Chowdhery, Google Brain* Developed a federated learning enabled custom deep learning model that powers VisionAir, an android application that predicts the Real-Time Air Quality Index of an image.	**Delhi, India**

Select Research Publications

Complete List at Google Scholar

[C] **MEGA: Multilingual Evaluation of Generative AI** [ ]
Kabir Ahuja, <u>Harshita Diddee</u>, ..., Kalika Bali, Sunayana Sitaram
To Appear in EMNLP 2023 **[EMNLP 2023]**

[C] **Fifty Shades of Bias: Normative Ratings of Gender Bias in GPT Generated English Text**
Rishav Hada, Agrima Seth, <u>Harshita Diddee</u>, Kalika Bali
To Appear in EMNLP 2023 **[EMNLP 2023]**

[C] **Too Brittle To Touch: Comparing the Stability of Quantization and Distillation Towards Developing Lightweight Low-Resource MT Models** [ ][Code]
<u>Harshita Diddee</u>, Sandipan Dandapat, Monojit Choudhury, Tanuja Ganu, Kalika Bali
Seventh Conference on Machine Translation **[WMT 2022]**

Select Research Projects

Interactive Neural Machine Translation-Lite (INMT-Lite) Jul'21 - May'23
Advisors: Dr. Monojit Choudhury, Dr. Tanuja Ganu, Dr. Sandipan Dandapat, Dr. Kalika Bali [**Code**]
> Built lightweight translation (<200MB) models for extremely low-resource languages like Gondi and Mundari (<25000 parallel sentences). Designed decoding pipeline to provide candidate translation suggestions to users.

Automatic Speech Recognition for Extremely Low-Resource Languages Oct'22 - Dec'22
Advisors: Dr. Sunayana Sitaram, Dr. Kalika Bali [ ][**Code**]
> Proposed the use of KenLM-based inference during training to select best-model more reliably.
> *Won third prize in The AmericasNLP Shared Task for Low-Resource ASR.*

VisionAir: Federated Learning Enabled Air Quality Estimation Jun'19 - Feb'20
Advisor: Dr.Aakanksha Chowdhery [ ][**Code**]
> Developed the compound deep neural network-based pipeline to replace the conventionally used convolution-based neural model so that we could train the model on edge..