Harshita **Diddee**

PhD Student, Carnegie Mellon University

Portfolio Github Google Scholar @ Email

Education

August 2023
Spring 2028
Carnegie Mellon University **Pittsburgh, USA**
PhD, Language Technologies Institute: Advised by Daphne Ippolito
Working on Data Curation for LLMs: (a) How can we quantify different dataset behaviors ? (b) Can we use these behaviors to discover modes of failure ? (c) and prune existing datasets ? (d) How can we compose existing datasets to create better datasets ?

May 2017
Jun 2021
Guru Gobind Singh Indraprastha University **Delhi, India**
B.Tech., Computer Science & Engineering | Department Rank: 2/120
Graduated as the Best Outgoing Student for the Class of 2021

Select Experience

Jul 2021
July 2023
Microsoft Research **Bangalore, India**
SCAI Centre Fellow | Primary Advisor: Dr. Kalika Bali, Microsoft Research India
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.

Jun 2022
Aug 2022
Frederick Jelinek Memorial Summer Workshop 2022 **Baltimore, USA**
Visting Pre-Doctoral Research | Host: Johns Hopkins University
Evaluated the generalizability of speech and text cross-lingual models to low-resource languages.

May 2019
Oct 2019
Indian Institute Of Technology, Delhi **Delhi, India**
Research Intern | Advisor: Aakanksha Chowdhery, Meta
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.

Select Research Publications

Complete List at Google Scholar

[R] **Chasing Random:Instruction Selection Strategies Fail to Generalize** [🔗]
Harshita Diddee, Daphne Ippolito
Under Review

[C] **Akal Badi ya Bias: An Exploratory Study of Gender Bias in Hindi Language Technology** [🔗]
Hada et. al.
Best Paper Award **[FAccT 2024]**

[C] **MEGA: Multilingual Evaluation of Generative AI** [🔗]
Kabir Ahuja, Harshita Diddee, ..., Kalika Bali, Sunayana Sitaram
EMNLP 2023 **[EMNLP 2023]**

Select Research Projects

Data Selection for Instruction Finetuning Feb'24 - Sep'24
[**Paper**]

- Demonstrated the brittle generalization of instruction selection strategies by focusing on if popular strategies cannot beat random baselines consistently.
- Exploring mitigating this brittleness using (a) methods that quantify the affordance (*How much variance can we get across a selection heuristic on a dataset) of datasets ?* and (b) designing an evaluation recipe that quantifies the impact on a model's learning ability (rather than its performance on a fixed benchmark) post training with selected data.

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**][**Paper**]

- 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. [**Paper**]

Select Research Projects

Automatic Speech Recognition for Extremely Low-Resource Languages Oct’22 - Dec’22

Advisors: Dr. Sunayana Sitaram, Dr. Kalika Bali [**Models**][**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 (Competition Track NeurIPS)*

VisionAir: Federated Learning Enabled Air Quality Estimation Jun’19 - Feb’20

Advisor: Dr.Aakanksha Chowdhery [🔗][**Code**]

- Created an air-pollution regression model that leveraged federated learning to train on user-contributed images of different environments mapped to different air pollution levels.
- 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..