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 designing LLM evaluation methods (a) for user-specific tasks at inference time (b) for diversity-seeking queries and (c) studying dataset behaviors for attributes for high-quality data.

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 speech translation and ASR for extremely low-resource languages.

May 2019 – Oct 2019 | **Indian Institute Of Technology, Delhi** | **Delhi, India**
Research Intern | Advisor: Aakanksha Chowdhery, Meta (now)
Developed a federated learning enabled custom deep learning model that powers that predicts the Air Quality Index of an image in real-time.

Select Research Publications

Complete List at Google Scholar

[C] Chasing Random:Instruction Selection Strategies Fail to Generalize [🔗][Code]
Harshita Diddee, Daphne Ippolito
To Appear in Findings of NAACL 2025 **[NAACL 2025]**

[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
Advisor: Dr. Daphne Ippolito [**Code**] [**Paper**]

- Demonstrated the brittle generalization of instruction selection strategies by showing that popular strategies cannot beat random baselines consistently.
- Showed that popular instruction following benchmarks have orthogonal performance trends while measuring performance of models on general instruction following capabilities which can hinder model selection.
- Developing a tool to quantify and improve the correlation between performance trends on these benchmarks for user-specific cases.

Automatic Speech Recognition for Extremely Low-Resource Languages Oct'22 - Jan'23
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)*

Select Research Projects

Interactive Neural Machine Translation-Lite (INMT-Lite) Jul'21 - Aug'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**]. Designing a constrained decoding pipeline to incorporate support for partial-input suggestion i.e., modelling suggestions on the partial input of a user.
- Designing automatic metrics to monitor *(a) what data is relevant to collect translations for ?* (b) what is the quality of submitted translation ? and (c) effort required by users to post-edit a noisy translation.

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..