Harshita **Diddee**

SCAI Center Fellow, Microsoft Research

🌐 Portfolio Github 🎓 Google Scholar @ Email

Education

May 2017
Jun 2021
Guru Gobind Singh Indraprastha University — **Delhi, India**
B.Tech., Computer Science & Engineering | Major GPA - 9.5/10, CGPA: 9.46/10
Department Rank: 2/120
Graduated as the Best Outgoing Student for the Class of 2021

Experience

Jul 2021
Present
Microsoft Research — **Bangalore, India**
SCAI Centre Fellow | Advisors: Dr. Kalika Bali
Developing edge friendly machine translation models for extremely low-resource languages. Exploring unsupervised methods of estimating data quality.

Jun 2022
Aug 2022
Johns Hopkins University — **Baltimore, USA**
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.

Oct 2020
Mar 2021
AI4Bharat — **Remote**
Research Intern | Advisor: Dr. Mitesh M. Khapra
Implemented a tesseract-based OCR pipeline; Assisted the mining of parallel sentences from a dense (12M+), monolingual embedding space (between a target and source language) using FAISS.

Mar 2020
Nov 2020
Indraprastha Institute of Information Technology Delhi (IIIT-D) — **Delhi, India**
Research Intern | Advisor: Dr.Koteswar Rao Jerripothula
Designed a Cross-Silo Federated Learning (FL) hypothesis for heterogeneous clients having mutually exclusive feature space.

May 2019
Oct 2019
Indian Institute Of Technology, Delhi (IIT-D) — **Delhi, India**
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.

Publications

S=In Submission, C=Conference, W=Workshop, J=Journal

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

[J] **CodeFed: Federated Learning enabled Code-Switching** [🔗]
Chetan Madan, Harshita Diddee, Deepika Kumar, Mamta Mittal
ACM Transactions on Asian and Low-Resource Language Information Processing **[TALLIP]**

[J] **Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages** [🔗][Code]
Ramesh et. al.
Transactions of the Association for Computational Linguistics **[TACL]**

[C] **The Six Conundrums of Building and Deploying Language Technologies for Social Good** [🔗]
Harshita Diddee*, Kalika Bali*, Monojit Choudhury*, Namrata Mukhija*
ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies **[ACM COMPASS]**

[W] **PsuedoProp at SemEval-2020 Task 11:Propaganda Span Detection using BERT-CRF and Ensemble Sentence Level Classifier** [🔗][Code]
Annirudha Chauhan, Harshita Diddee
Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval) **[SemEval]**

[C] **CrossPriv: User privacy preservation model for cross-silo federated software** [🔗] [Code]
Harshita Diddee, Bhrigu Kansra
35th IEEE/ACM International Conference on Automated Software Engineering (ASE) **[ASE]**

Select Research Projects

Interactive Neural Machine Translation-Lite (INMT-Lite) Jul’21 - Present

Advisors: Dr. Tanuja Ganu, Dr. Monojit Choudhury, Dr. Sandipan Dandapat, Dr. Kalika Bali [**Code**]

- Prototyping with extremely low-resource languages like Gondi and Mundari.
- Evaluating if a low accuracy translation model can assist users by providing candidate translations. Designing metrics to monitor the users effort and the quality of the incoming data.
- Identifying what would be the best interface (dropdown lists, Bag of Words, gisting, etc) to present such low-quality assistance so as to not make it disruptive for the user’s task to providing data. Exploring a method of constrained decoding to include partial-input assistance.

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

Advisors: Dr. Sunayana Sitaram, Dr. Kalika Bali [🔗][**Code**]

- Exploring mechanisms of leveraging text-based language models in improving model selection for ASR models.
- Ranked 3rd at the AmericasNLP Shared Task for Low-Resource ASR.

VisionAir Jun’19 - Feb’20

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

- I developed the compound deep neural network-based pipeline to replace the conventionally used convolution-based neural model. This alternative model effectively reduced VisionAir’s model size by nearly 400 times, making it accessible to even those who did not own mobile devices capable of intensive computation.
- Work published by TensorFlow

Academic Service

Workshop Co-Organizer SLT’22 Hackathon at IEEE Spoken Language Technology Workshop

Peer Reviewer SLT’22

Honours and Grants

ACM Grant to present work at the 35th IEEE/ACM International Conference on ASE

Google LLC Travel Grant to attend the TensorFlow Dev Summit at Sunnyvale, CA

The Marconi Society, Google LLC Runner’s Up at the Celestini Prize India 2019

Government of Singapore and India Runner’ Up at the Singapore India Hackathon

Government of India Winner for e-Yantra National Robotics Competition

Google AI Selected to attend the Google AI Summer School

Volunteering Roles

TEM Reading Group, MSR India *Organizer* 2022 - Present

- Organizer for Technology and Empowerment Reading Group at Microsoft Research India

Volunteer and Speaker at PyData, NumFOCUS *Volunteer | Speaker* 2019 - 2021

- Talk on Open Problems in Federated Learning [🌐]

Invited Talk, Women In Data Science *Speaker*

- Talk on Emerging Setups in Federated Learning [🌐]

References

- Dr. Kalika Bali .... *Principal Researcher, Microsoft Research, India* [🌐]
- Dr. Mitesh M. Khapra .... *Associate Professor, IIT Madras, India* [🌐]
- Dr. Monojit Choudhury .... *Principal Data and Applied Scientist, Microsoft Turing, India* [🌐]
- Dr. Aakanksha Chowdhery .... *Staff Software Engineer, Google* [🌐]