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 furthering personalization through (a) evaluating data curation for tasks that align with specific/nuanced user needs (b) adapting LLM reasoning patterns to user profiles and (c) exploring how agentic populations can emulate human populations at scale.

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

May 2025 – Aug 2025 | **Amazon Core Search** | **Palo Alto, CA**
Applied Science Intern| Primary Advisor: Dr.Tanya Roosta, Amazon Science
Built an end-end simulation environment for estimating the impact of treatments on Amazon Search Customers using simulated AI Agents. Conducted a click-tendency evalutation to show how LLMs drift from exact matches of user intents, hallucinate rationales, and over-permit low-quality items under certain conditions.

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.

Select Research Publications

Complete List at Google Scholar

[C] NoveltyBench: Evaluating Language Models for Humanlike Diversity [🔗][Code]
Yiming Zhang, Harshita Diddee ..., Daphne Ippolito
Proceedings of COLM 2025 [**COLM 2025**]

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

Select Research Projects

Retrieval Tool for Selecting the Right Benchmark Jan'25 - Nov'25
Advisor: Dr. Daphne Ippolito [**Website**]

- Built a retriever that surfaces benchmark items matching a practitioner's use-case with high precision (Human Evaluation with 11 ML practitioners).
- Show that this tool can help diagnose low content validity i.e., coverage gaps across close variants of the same task (Eg: Testing models excessively on formatting for json but not YAML).
- and poor convergence validity i.e., getting divergent model rankings across "same-capability" benchmarks (Eg: Getting 2 different best models among n candidates when 2 'reasoning' benchmarks are used).

Evaluating Robustness of Post Training Data Selection 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.