Anjali Kantharuban

anjaliruban@cmu.edu • github.com/AnjaliRuban • anjaliruban.com

Education

Carnegie Mellon University August 2023 - Present

PhD Language and Information Technology

University of Cambridge October 2022 - June 2023

MPhil Theoretical and Applied Linguistics by Thesis

University of California, Berkeley August 2018 - May 2022

BA Linguistics & BA Computer Science with Honors GPA: 3.96

Research Experience

Cambridge Language Technology Laboratory October 2022 - June 2023

Research Assistant *Advised by Anna Korhonen*

- Involved in preparing a grant proposal for the Cambridge-LMU Strategic Partnership for 40,000 euros for the purpose of dialectal data collection.
- Analyzed the performance gap between regional dialects of various languages across machine translation and automatic speech recognition. [2]

Berkeley Natural Language Processing Group March 2020 - June 2022

Undergraduate Researcher *Advised by Dan Klein and Trevor Darrell*

- Used PyTorch to implement networks for classification, generation, and embedding tasks over datasets such as ALFRED and BabyAI.
- Developed a novel data augmentation technique for adapting imitation learning instruction following agents to out-of-domain instructions, primarily those which explicitly contain multiple goals.

MIT Language & Intelligence Lab June 2021 - June 2022

Undergraduate Researcher (REU) *Advised by Jacob Andreas*

- Expanded work regarding lexicon learning for sequence modeling to low resource languages with structural differences from English. [1]

Publications

[1] **Pangea: A Fully Open Multilingual Multimodal LLM for 39 Languages**

Xiang Yue*, Yueqi Song*, Akari Asai, Simran Khanuja, Anjali Kantharuban, Seungone Kim, Jean de Dieu Nyandwi, Lintang Sutawika, Sathyanarayanan Ramamoorthy, Graham Neubig

Under Review at ICLR 2025

[2] **Stereotype or Personalization? User Identity Biases Chatbot Recommendations**

Anjali Kantharuban*, Jeremiah Milbauer*, Emma Strubell, Graham Neubig

Under Review at TACL

[3] **Quantifying the Dialect Gap in Large Language Models and its Causes Across Languages**

Anjali Kantharuban, Ivan Vulić, Anna Korhonen

Findings of EMNLP, 2023.

[4] **Counting the Bugs in ChatGPT's Wugs: A Multilingual Investigation into the Morphological Capabilities of a Large Language Model**

Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schuetze, Kemal Oflazer, David R Mortensen

EMNLP, 2023.

[5] **One-Shot Lexicon Learning for Low-Resource Machine Translation**

Anjali Kantharuban, Jacob Andreas

Presented at Widening Natural Language Processing Workshop, EMNLP, 2021.

Teaching Experience

Subword Modeling at Carnegie Mellon University January 2024 - May 2024

Data Structures, CS61B at UC Berkeley August 2019 - May 2022
Head Teaching Assistant

Awards & Scholarships

NSF Graduate Research Fellowship Fall 2023 - Present
Gates Cambridge Scholarship Fall 2022 - Spring 2023
Fulbright US-UK Student Program (Alternate) Fall 2022
Outstanding Graduate Student Instructor Award Spring 2022
Foreign Language and Area Studies Fellowship Fall 2020 - Spring 2021