Abhinav Rao

https://abhinavrao.netlify.app
Master's @ Carnegie Mellon University

☎ +1-(412)-251-8197
✉ abhinavr@andrew.cmu.edu, abhinav.797c@gmail.com
https://linkedin.com/in/abhinav-rao

EDUCATION

Carnegie Mellon University, School of Computer Science — Pittsburgh, PA
Master of Science (NLP), Language Technology Institute — Expected Dec '24, GPA: 4.2/4
Coursework: Multimodal Machine Learning, Advanced Natural Language Processing, Quantitative Evaluation of language technologies

Birla Institute of Technology and Science (BITS), Pilani — Hyderabad, India
Bachelor of Engineering in Computer Science — Feb '22, GPA: 9.26 / 10.0
Coursework: Machine Learning, Data Structures and Algorithms, Software Engineering, Compilers

SELECT PUBLICATIONS

[1] **Abhinav Rao**, Sachin Vashistha*, Atharva Naik*, Somak Aditya, and Monojit Choudhury. Tricking llms into disobedience: Understanding, analyzing, and preventing jailbreaks. LREC-CoLING 2024, abs/2305.14965, 2023.

[2] **Abhinav Rao**, Thi-Nga Ho, and Eng-Siong Chng. Punctuation restoration for singaporean spoken languages. Asia-Pacific Speech and Information Processing Association, 2022.

[3] **Abhinav Rao***, Aditi Khandelwal*, Kumar Tanmay*, Utkarsh Agarwal*, and Monojit Choudhury. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in llms. EMNLP 2023, 2023.findings-emnlp.892, 2023.

[4] Aashiq Muhamed*, Harshita Diddee*, and **Abhinav Rao***. Less is fed more: Sparsity mitigates feature distortion in federated learning. MOOMIN, EACL, 2024.

EXPERIENCE

Bell Labs — Murray Hill, NJ
Research Intern, Autonomous Systems — June 2024 - Present

- Developing LLM agents with AutoGen and LangChain for Code-repair.

Microsoft — Bangalore, India
AI Resident (Advisor: Prof. Monojit Choudhury) — Aug 2022 - Jul 2023

- Improved Bing Chat classifier performance by **5% and 17%** (F1-score), for jailbreaking and content-harm through offline data curation.
- Evaluated jailbreak effectiveness against 9 different Large Language Models (LLMs) by formalizing LLM jailbreaking and proposing a new evaluation criteria. Accepted @ LREC-CoLING 2024. Code Available.
- Determined a western-centric ethical-reasoning bias in language models and posited a framework for fairness in LLMs. Published at EMNLP Findings '23.

Research Intern (Microsoft Research) (Advisor: Dr. Sunayana Sitaram) — Jan 2022 - Jul 2022

- Sped up Bing query expansion by **98%** through multilingual data augmentation using vectorDB (DiskANN) and topic modeling (`BERTopic`).

Nanyang Technological University (NTU) — Singapore
Research Intern (SpeechLab, NTU. Advisor: Prof. Chng Eng Siong) — Jun 2021 - Dec 2021

- Beat the SOTA Chinese Model by **4.2%** F1-score on automatic punctuation restoration of Automatic Speech Recognition (ASR) text. Extended punctuation to Malay with XLM-R through a pretraining-style objective Published at APSIPA'22. Code available.

Oracle Corporation — Bangalore, India
Software Developer Intern — Jun 2021 - Jul 2021

- Achieved a **75%** accuracy on automatic bug categorization through text-rank and word-embedding matching on very-large databases (VLDB) using `Oracle-CX` and `nltk`.

SKILLS

Programming Languages & Libraries: Python; C++; SQL; NumPy, SciPy; Pandas; PyTorch; Tensorflow; Keras; Scikit-learn;
General skills: Machine Learning (ML); Natural Language Processing (NLP); Deep Learning (DL); LLMs

SELECT PROJECTS

Compositionality of Vision Language Models — Carnegie Mellon University
Course project for Multimodal Machine Learning — September 2022 - Dec 2022

- Achieved a **5%** improvement on Winoground by finetuning BLIP-2 with a multi-objective loss using counterfactual text and image generation. Involved dependency parsing, object segmentation, and image-inpainting in the generation pipeline Report. Code Available.

Multilingual Sparse Federated Learning — Carnegie Mellon University
Course project for Advanced Natural Language Processing — September 2022 - Present

- Analyzed the impact of parameter efficient training methods for Machine translation in a federated learning setting.
- Quantified the tradeoffs across high and low resource languages, showing robustness to the curse of multilinguality. Accepted at MOOMIN, EACL'24.