

Abhinav Rao

<https://aetherprior.github.io>
Master's student @ Carnegie Mellon University

+1-(412)-251-8197
✉ abhinav.797c@gmail.com
🐙 [GitHub Profile](#)
🌐 [LinkedIn Profile](#)

EDUCATION

Carnegie Mellon University, School of Computer Science

Master of Science in Intelligent Information Systems
Coursework (Ongoing): Multimodal Machine Learning, Advanced Natural Language Processing

Pittsburgh, PA
Expected Dec '24

Birla Institute of Technology and Science (BITS), Pilani

B.Engg. in Computer Science
Coursework: Machine Learning, Data Structures and Algorithms, Software Engineering, Compilers

Hyderabad, India
Feb '22, GPA: 9.26 / 10.0

EXPERIENCE

Microsoft Turing

Research Fellow

Bangalore, India
Aug 2022 - Jul 2023

- Curated datasets for Content-harm and Jailbreaks on Bing Chat to improve classifier performance by 5% and 17% (F1-score), for jailbreak and content-harm detection respectively.
- Formalized and studied the jailbreak phenomenon. Ideated a taxonomy of jailbreaks and evaluated their effectiveness against different GPT-based Large Language Models (LLMs). Submitted for publication to EMNLP.
- Evaluated ethical generalization capabilities of different LLMs. Submitted for publication to EMNLP.

Microsoft Research

Research Intern

Bangalore, India
Jan 2022 - Jul 2022

- Designed a multilingual data augmentation tool for query expansion as part of Project LITMUS. Sped up the pipeline for Bing's Defensive team by 10x using a multilingual topic model and Approximate Nearest Neighbors (ANNS).

Nanyang Technological University (NTU)

Research Intern

Singapore
Jun 2021 - Dec 2021

- Researched and developed a deep learning model for multilingual automatic punctuation restoration of Automatic Speech Recognition (ASR) text as part of SpeechLab, NTU. Beat the best-performing Chinese Model by 4.2% F1-score. Published at APSIPA'22.

Oracle Corporation

Software Developer Intern

Bangalore, India
Jun 2021 - Jul 2021

- Engineered a system to map bugs to their associated features, incorporating text-mining from large databases for bug-feature-customer analytics. Achieved a 75% accuracy on bug analysis. Used Python and libraries such as NLTK and SpaCy.

PROJECTS

Auto Code Commenting

github.com/AetherPrior/Code-commenting

BITS Pilani, Hyderabad
Jan 2021 - Dec 2021

- Enhanced existing code-commenting models, using an encoder-decoder architecture with Pointer Generator networks and coverage attention. Achieved a BLEU-4 score of 40.

Timetable Helper

github.com/crux-bphc/Chronofactorem

BITS Pilani, Hyderabad
Jan 2020 - May 2020

- Developed a web application for helping students with registration by analyzing course statistics to project demand.
- Achieved a record 1500 user-registrations within the first day of enrollment. Deployed and maintained by the current students of BITS Pilani, Hyderabad.

SKILLS

Programming Languages and Libraries: Python; C++; NumPy, SciPy & Pandas; PyTorch; HuggingFace; NLTK & SpaCy
General skills: Machine Learning; Natural Language Processing

PUBLICATIONS

- [1] A. Rao, T.-N. Ho, and E.-S. Chng. Punctuation restoration for singaporean spoken languages. *Asia-Pacific Speech and Information Processing Association*, 2022.
- [2] A. Rao, S. Vashistha, A. Naik, S. Aditya, and M. Choudhury. Tricking LLMs into disobedience: Understanding, analyzing, and preventing jailbreaks. *ArXiv*, abs/2305.14965, 2023.