Abhinav Rao

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

☎ +1-(412)-251-8197
✉ abhinavr@andrew.cmu.edu
GitHub Profile
LinkedIn Profile

EDUCATION

Carnegie Mellon University, School of Computer Science — Pittsburgh, PA
Master of Science in Intelligent Information Systems — Expected Dec '24
Coursework (Ongoing): Multimodal Machine Learning, Advanced Natural Language Processing

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

EXPERIENCE

Microsoft Turing — Bangalore, India

Research Fellow — 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.
- Posited a moral-alignment framework for cross-cultural fairness in LLMs, and evaluated their ethical reasoning capabilities. Paper accepted at EMNLP Findings 2023. Preprint on arxiv
- Formalized and studied the jailbreak phenomenon. Ideated a taxonomy of jailbreaks and evaluated their effectiveness against different GPT-based Large Language Models (LLMs). In review at LREC-CoLING 2024. Preprint on arxiv

Microsoft Research — Bangalore, India

Research Intern — 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) — Singapore

Research Intern — Jun 2021 - Dec 2021

- Researched and developed a deep learning BERT-based 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 — Bangalore, India

Software Developer Intern — 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 — BITS Pilani, Hyderabad

github.com/AetherPrior/Code-commenting — 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 — BITS Pilani, Hyderabad

github.com/crux-bphc/Chronofactorem — 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 Langauges and Libraries: Python; C++; SQL; NumPy, SciPy; Pandas; PyTorch; Tensorflow; Keras; Scikit-learn; HuggingFace; NLTK; SpaCy
General skills: Machine Learning (ML); Natural Language Processing (NLP); Deep Learning (DL); Neural networks; Research; Foundation models; GenAI; Software Engineering; Linux; Data Science; Data Analytics;

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, A. Khandelwal, K. Tanmay, U. Agarwal, and M. Choudhury. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in llms, 2023.

[3] 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.