Abhinav Rao

+1-(412)-251-8197
abhinav.797c@gmail.com
GitHub Profile
LinkedIn Profile

Graduate student with 1.5 years of experience in Natural Language Processing.
Looking for summer internships in Data Science and Machine Learning.

EDUCATION

Carnegie Mellon University | M.S. in Intelligent Information Systems, School of Computer Science
Expected Dec '24, Pittsburgh, PA

Birla Institute of Technology and Science, Pilani | B.Engg. in Computer Science, GPA: 9.26 / 10.0
Feb '22, Hyderabad, India

EXPERIENCE

Microsoft Turing | Research Fellow | Aug 2022 - Jul 2023 | Location: Bangalore, India

- 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 | Jan 2022 - Jul 2022 | Location: Bangalore, India

- 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 | Jun 2021 - Dec 2021 | Location: Singapore

- 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 | Intern | Jun 2021 - Jul 2021 | Location: Bangalore, India

- 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 | Jan 2021 - Dec 2021 | github.com/AetherPrior/Code-commenting

- Enhance 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 | Jan 2020 - May 2020 | github.com/crux-bphc/Chronofactorem

- 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.

COURSEWORK

• Machine Learning • Deep Learning •Data Structures and Algorithms • Software Engineering • Compilers

SKILLS

• Python • C++ • NumPy, SciPy & Pandas • PyTorch • Natural Language Processing • HuggingFace • NLTK & SpaCy

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.