Abhinav Rao

Pittsburgh, PA (412) 251-8197 abhinavr@cs.cmu.edu linkedin.com/in/abhinav-rao
abhinavrao.netlify.app github.com/Aetherprior scholar.google.com/citations?user=U_wk4ssAAAAJ

Education

Carnegie Mellon University - Language Technologies Institute, School of Computer Science *Dec 2024*
Master of Science in Natural Language Processing (Artificial Intelligence)*; **GPA**: 4.2/4* Pittsburgh, PA

Coursework: Advanced Natural Language Processing (A+), Multimodal Machine Learning (A+), Quantitative Evaluation of Language Technologies (A), Ethics in AI (A+)

Birla Institute of Technology and Science, Pilani - Department of Computer Science *Feb 2022*
Bachelor of Engineering in Computer Science*; **GPA**: 9.26/10* Hyderabad, India

Select Publications

[1] **Abhinav Rao**, Sachin Vashistha*, Atharva Naik*, Somak Aditya and Monojit Choudhury. Tricking LLMs into Disobedience: Understanding, Analyzing, and Preventing Jailbreaks *in LREC-CoLING 2024*

[2] **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 *in EMNLP 2023*

Experience

Bell Labs *June 2024 – Present*
Research Intern, Autonomous Systems Murray Hill, NJ

- Developing Large Language Model (LLM) agents with AutoGen and LangChain for Code-repair.

Microsoft *Jan 2022 – Jul 2023*
AI Resident (Turing Team) (Advisor: Prof. Monojit Choudhury) Bangalore, India

- 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 LLMs by formalizing LLM jailbreaking. Showed an inverse scaling trend: GPT3.5 is 20% more susceptible than FLAN-T5.
- Determined western-centric reasoning biases in language models; posited a framework to enhance fairness in LLMs.

Research Intern (Microsoft Research) (Advisor: Dr. Sunayana Sitaram)

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

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

- Beat SOTA by 4.2% F1-score on Chinese punctuation restoration of ASR text, with a pretraining-style objective. Extended punctuation to English and Malay with XLM-R. **Accepted at APSIPA'22**

Oracle *Jun 2021 – Jul 2021*
Software Developer Intern Bangalore, India

- 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.

Projects

Compositionality of Vision Language Models *Sep 2022 - Dec 2022*

- Achieved 5% improvement on Winoground by finetuning BLIP-2 using counterfactual text and image generation.
- Involved dependency parsing, object segmentation, and image-inpainting in the generation pipeline.

Multilingual Sparse Federated Learning *Sep 2022 - Present*

- Analyzed parameter-efficient finetuning (PEFT) for machine translation in federated learning, quantifying tradeoffs across languages and demonstrating robustness to multilinguality; **research accepted at MOOMIN, EACL'24.**

Skills

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