

Abhinav Rao

Master's Student, Language Technologies Institute, CMU

🌐 abhinavrao.netlify.app @ abhinavr@andrew.cmu.edu 📄 github.com/aetherprior 🎓 Google Scholar
📍 Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, 15217

Education

Aug 2023 Dec 2024	Carnegie Mellon University Language Technologies Institute M.S. student in Intelligent Information Systems (MIIS)	Pittsburgh, PA
Feb 2022 Aug 2018	Birla Institute of Technology and Science (BITS) Pilani B.E. Computer Science (Graduated Early)	Hyderabad, India

Select Experience

Aug 2024 Jun 2024	Bell Labs [🔗] <i>Research Intern, Autonomous Systems</i> Constructed a code repair prototype using multi-agent pipeline with Large Language Models (LLMs).	Murray Hill, NJ
Jul 2023 Aug 2022	Microsoft [🔗] <i>Research Fellow (Turing Team)</i> Worked on Responsible AI (RAI) focusing on AI Ethics and Safety. Analyzed ethical reasoning capabilities of LLMs, and their susceptibility to jailbreaks.	Bangalore, India
Jul 2022 Jan 2022	Microsoft Research [🔗] <i>Research Intern</i> Developed a multilingual query expansion tool with embedding interpolation and topic modeling.	Bangalore, India
Dec 2021 Jun 2021	Nanyang Technological University, SpeechLab [🔗] <i>Research Intern (SpeechLab)</i> Extended punctuation restoration capabilities to Chinese and Malay with XLM-R . Improved F1-score by 4.2% over state-of-the-art for Chinese punctuation restoration in ASR text using a pretraining-style objective.	Singapore

Publications

S=In Submission, C=Conference, W=Workshop, P=Preprint

- [C.1] **Tricking LLMs into Disobedience: Understanding, Analyzing, and Preventing Jailbreaks** [🔗]
[Abhinav Rao](#), Sachin Vashistha*, Atharva Naik*, Somak Aditya, and Monojit Choudhury
[Published at LREC-CoLING 2024]
- [C.2] **Ethical Reasoning over Moral Alignment: A Case and Framework for In-Context Ethical Policies in LLMs** [🔗]
[Abhinav Rao*](#), Aditi Khandelwal*, Kumar Tanmay*, Utkarsh Agarwal*, Monojit Choudhury
[Published at the Findings of EMNLP 2023, Presented as a Keynote at WinLP]
- [S.1] **Normad: A benchmark for measuring the cultural adaptability of large language models** [🔗]
[Abhinav Rao*](#), Akhila Yerukola*, Vishwa Shah, Katharina Reinecke, and Maarten Sap
[Accepted Non-archivally @ C3NLP, ACL 2024, Under Review at NAACL 2025]
- [C.3] **Punctuation Restoration for Singaporean Spoken Languages** [🔗]
[Abhinav Rao](#), Thi-Nga Ho, and Eng-Siong Chng
[Asia-Pacific Speech and Information Processing Association 2022]
- [W.1] **Less is Fed More: Sparsity Mitigates Feature Distortion in Federated Learning** [🔗]
Aashiq Muhamed*, Harshita Diddee*, [Abhinav Rao*](#)
[CustomNLP4U, EMNLP 2024, Also Presented at MOOMIN, EACL 2024]
- [P.1] **Jailbreak Paradox: The Achilles' Heel of LLMs** [🔗]
[Abhinav Rao*](#), Monojit Choudhury*, and Somak Aditya*
[arXiv preprint arXiv:2406.12702]
- [S.2] **MALITE: Lightweight Malware Detection and Classification for Constrained Devices** [🔗]
Siddharth Anand, Barsha Mitra, Soumyadeep Dey, [Abhinav Rao](#), Rupsa Dhar, and Jaideep Vaidya
[arXiv preprint arXiv:2309.03294, [Under review at IEEE-TETC]]

Select Research Projects

Jailbreaking Language Models

Nov '22 - Present

Advisors: [Prof. Monojit Choudhury](#), [Prof. Aditya Somak](#)

- Evaluated jailbreak effectiveness against 9 different LLMs by formalizing LLM jailbreaking, showing an inverse scaling trend where GPT-3.5 is 20% more susceptible than FLAN-T5. [🔗] [Lrec-CoLING'24] (Coverage: [TCS Research Webinar on Genrative AI](#)).
- Developing a theoretical framework to explain the jailbreak-paradox, explaining the inverse scaling phenomenon in toxicity/jailbreaking. (Work-in-progress covered by [Analytics IndiaMag](#)).
- Improved **Bing Chat** classifier performance by 5% and 17% (F1-score) for jailbreaking and content-harm detection through offline data curation.

Ethical Reasoning Capabilities of LLMs

Aug'22 - July'23

Advisors: [Dr. Monojit Choudhury](#)

- Designed a framework to evaluate the ethical reasoning capabilities of Language models over increasing granularities of ethical policies. Uncovered a bias favoring western centric ethical principles in GPT-4. [EMNLP Findings '23] [Keynote at WiNLP '23]

Cultural Reasoning of LLMs

Sep'23 - Oct'24

Advisors: [Prof. Maarten Sap](#), [Prof. Katharina Reinecke](#)

- Built a benchmark dataset of 2.6k cultural situations spanning 75 countries measuring cultural biases in LLMs
- Measured cultural adaptability of 17 language models, determining strong sycophancy and western-centric biases. [Under review at NAACL '25] [Presented at C3NLP, ACL '24]

Multilingual Federated Learning

Sep 23' - April 24'

Independent Research

- Compared and contrasted different parameter-efficient finetuning (PEFT) techniques, such as sparse subnets and LoRA for machine translation in federated learning [Presented at MOOMIN, EACL '24] [Accepted at CustomNLP4U, EMNLP '24]

Talks

“Less is Fed More: Sparsity Mitigates Feature Distortion in Federated Learning”

- MOOMIN, EACL '24, Malta [🔗] [📺]

March 2024 (Remote)

“Punctuation Restoration for Singaporean Spoken Languages”

- APSIPA '22, Chiang-Mai, Thailand [🔗] [📺]

November 2022 (Remote)

Honours and Awards

Amazon Trusted AI Challenge Grant, '24 [🔗] Awarded \$250,000 as a model developer team for the Amazon Trusted AI challenge.

BITS Merit Scholarship, '18, '22 Tuition waiver of \$3300 (INR 280,000 total) awarded to the top 3%ile of students for academic excellence.

Teaching

Advanced Natural Language Processing (CMU-LTI 11711) Teaching Assistant

Sep'24 - Dec'24

- Responsibilities included conducting tutorials, evaluating assignments, and helping students with the assignments and advising them on their course projects.

Academic Service

Reviewer	ACL ARR Dec '23, TPAMI '24
Sub-Reviewer	NAACL'22
Volunteer	Panini Linguistics Olympiad (PLO) '23

References

- Prof. Maarten Sap Assistant Professor, Carnegie Mellon University [🔗]
- Prof. Monojit Choudhury Professor, MBZUAI, UAE [🔗]
- Prof. Somak Aditya Assistant Professor, IIT-KGP, India [🔗]
- Dr. Sunayana Sitaram Principal Researcher, Microsoft Research, India [🔗]