

Chahat Raj

PhD in Computer Science | George Mason University

🌐 chahatraj.github.io @ craj@gmu.edu 🌐 github.com/chahatraj 🎓 Google Scholar

Research Interests

I work in Responsible AI, Ethics, and Fairness, specifically in the evaluation and mitigation of socio-cultural biases within multilingual and multimodal NLP applications with a focus on LLMs. I have recently published at EMNLP, AIES, and ECIR.

Education

Present Aug 2022	George Mason University Ph.D. in Computer Science Advisors: Ziwei Zhu , Antonios Anastasopoulos	Fairfax, USA
Aug 2021 Aug 2019	Delhi Technological University Masters in Information Systems (Research Track) Advisor: Priyanka Meel	Delhi, India
May 2019 Aug 2015	Indira Gandhi Delhi Technical University for Women Bachelors in Computer Science & Engineering	Delhi, India

Experience

Present May 2024	George Mason University School of Computer Science <i>Graduate Research Assistant</i> Advisors: Ziwei Zhu , Antonios Anastasopoulos <u>Projects</u> : Identifying Attribution Biases in Large Language Models; Exploring Stereotypical Associations in Large Vision-Language Models through generative tasks - image generation and question answering.	Fairfax, VA
Sep 2024 May 2024	University of Washington The Information School <i>Visiting Researcher</i> Advisor: Aylin Caliskan <u>Project</u> : Social Biases in Visual Question Answering in Generative AI.	Seattle, WA
Aug 2023 May 2023	George Mason University School of Computer Science <i>Graduate Research Assistant</i> Advisor: Ziwei Zhu <u>Project</u> : Fairness and Interpretability of Transformer Models.	Fairfax, VA
Jul 2022 Jun 2021	University of Technology Sydney School of Computer Science <i>Visiting Scholar</i> Advisor: Mukesh Prasad <u>Projects</u> : Disaster Fundraising on Social Media Analysis with the Australian Red Cross, AI integration in aboriginal and indigenous communities' lifestyle, Web information pollution detection (cyberbullying, hate, and offensive speech), Machine learning-based decision-making in clinical vertigo diagnosis.	Remote
Aug 2021 May 2021	Indian Institute of Management Raipur <i>Researcher</i> Advisor: Manojit Chattopadhyay <u>Projects</u> : Cryptocurrencies' price prediction, Analysing causal relationships between stock market behavior and societal events.	Remote
Aug 2022 Aug 2019	Delhi Technological University <i>Graduate Student Researcher</i> Advisor: Priyanka Meel <u>Projects</u> : Developing deep learning frameworks to detect antisocial web content (fake news, infodemic, misinformation, disinformation, rumor, death hoax, and clickbait), Infodemic impact analysis.	Delhi, India

Selected Publications

[full list here](#)

- [C.5] **BiasDora: Exploring Hidden Biased Associations in Vision-Language Models** [PDF | Code]
[Chahat Raj](#), Anjishnu Mukherjee, Aylin Caliskan, Antonios Anastasopoulos, Ziwei Zhu
Conference on Empirical Methods in Natural Language Processing [EMNLP Findings'24]
- [C.4] **Breaking Bias, Building Bridges: Evaluation and Mitigation of Social Biases in LLMs via Contact Hypothesis** [PDF | Code]
[Chahat Raj](#), Anjishnu Mukherjee, Aylin Caliskan, Antonios Anastasopoulos, Ziwei Zhu
AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society [AIES'24]

- [C.3] **SALSA: Saliency-Based Switching Attack for Adversarial Perturbations in Fake News Detection Models** [PDF | Code]
 Chahat Raj*, Anjishnu Mukherjee*, Hemant Purohit, Antonios Anastasopoulos, Ziwei Zhu
European Conference on Information Retrieval [ECIR'24]
- [C.2] **Global Voices, Local Biases: Socio-cultural Prejudices across Languages** [PDF | Code]
 Anjishnu Mukherjee*, Chahat Raj*, Ziwei Zhu, Antonios Anastasopoulos
Conference on Empirical Methods in Natural Language Processing [EMNLP'23]
- [C.1] **True and Fair: Robust and Unbiased Fake News Detection via Interpretable Machine Learning** [PDF | Code]
 Chahat Raj, Anjishnu Mukherjee, Ziwei Zhu
AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society [AIES'23]

Skills

Natural Language Processing, Generative AI, Multimodal Large Language Models, Model Training, Evaluation & Validation, Instruction-tuning, Prompt Engineering, LLM-human alignment, Vision, Ethics and Bias Mitigation, Explainability.

Talks

“Breaking Bias, Building Bridges”

› Talk at AAAI/ACM AIES October 2024 (San Jose, USA)

“LLM Ethics - Bias and Mitigation”

› Invited Lecture - CS 678 Advanced NLP @ George Mason University October 2024 (Virginia, USA)

“A Psychological View to Social Biases in LLMs”

› SouthNLP @ Emory University April 2024 (Georgia, USA)

Awards

OpenAI Researcher Access Program Award 2025 Funded \$1000 worth of OpenAI API credits.

Travel Grant by AAAI/ACM AIES 2024 & 2023 Funded \$2000 to attend AIES in San Jose, USA and Montreal, Canada.

CAHMP Fellowship @ George Mason University, 2024 Summer research on social biases in question answering.

Best Paper Award @ MASC-SLL 2024, Johns Hopkins University For the paper ‘A Psychological View to Social Bias in LLMs: Evaluation and Mitigation’.

Summer Graduate Research Award @ George Mason University, 2023 For research on ‘Explainability and Robustness of Debiasing Approaches for Transformers’.

Research Excellence Award @ Delhi Technological University 2023 (×2) & 2022 Received three Commendable Research Awards with a total cash prize of INR 150k by DTU, Delhi, for the research published in reputed scientific journals.

Graduate Scholarship @ Delhi Technological University 2019 Received INR 297k by AICTE for qualifying GATE.

Teaching

CS 108 Introduction to Computer Programming, GMU *Teaching Assistant* Spring'24

› Responsible for teaching labs, one-to-one student tutoring, creating lab modules, grading, and exam invigilation.

CS 478 Natural Language Processing, GMU *Teaching Assistant* Fall'23

› Delivered ad-hoc lectures, student tutoring, teaching labs, and assignment grading.

COMP 502 Mathematical Foundations of Computing, GMU *Teaching Assistant* Fall'23

› Responsible for one-to-one student tutoring sessions, and assignment grading.

CS 112 Introduction to Python Programming, GMU *Teaching Assistant* Fall'22, Spring'23

› Responsible for labs, one-to-one student tutoring, creating lab modules, assignment grading, and exam invigilation.

Academic Service

Reviewer ACM Transactions on Intelligent Systems and Technology 2024, LTEDI @ EACL 2024, TrustNLP 2024, CSCW 2024, NeurIPS 2023, Elsevier & Springer Journal papers

Sub-Reviewer EMNLP'24, AIES'24

Student Mentorship

› Mamnuya Rinki [MS, GMU]