

# Chahat Raj

## PhD in Computer Science | George Mason University

🌐 [chahatraj.github.io](https://chahatraj.github.io) @ [craj@gmu.edu](mailto:craj@gmu.edu) 📄 [Google Scholar](https://scholar.google.com/citations?user=ChahatRaj) 📁 [github.com/chahatraj](https://github.com/chahatraj) 🐦 [chahatsaidit](https://twitter.com/chahatsaidit)

• NLP • LLMs & VLMs • LLM Agents • Multimodality • Multilinguality • Alignment • Responsible AI • Ethics • Cognitive Sciences

### Research Interests

I work on NLP and Generative AI, specializing in Large Vision-Language Models, focusing on their harm evaluation & mitigation, interpretability, and alignment. My research explores socio-cultural biases, ethics, and fairness in multilingual and multimodal LLMs, aiming for responsible AI development. I have recently published at EMNLP, AIES, and ECIR.

### Education

<b>Present</b> <b>Aug 2022</b>	<b>George Mason University</b> Ph.D. in Computer Science (GPA: 3.83) Advisors: <a href="#">Ziwei Zhu</a> , <a href="#">Antonios Anastasopoulos</a>	<b>Virginia, USA</b>
<b>Aug 2021</b> <b>Aug 2019</b>	<b>Delhi Technological University</b> Masters in Information Systems (Research Track) Advisor: <a href="#">Priyanka Meel</a>	<b>Delhi, India</b>
<b>May 2019</b> <b>Aug 2015</b>	<b>Indira Gandhi Delhi Technical University for Women</b> Bachelors in Computer Science & Engineering	<b>Delhi, India</b>

### Experience

<b>Present</b> <b>May 2024</b>	<b>University of Washington, George Mason University</b> <i>Graduate Research Assistant</i>   Advisors: <a href="#">Ziwei Zhu</a> , <a href="#">Antonios Anastasopoulos</a> , <a href="#">Aylin Caliskan</a> <ul style="list-style-type: none"><li>Simulating societal dynamics using LLMs to audit harms and biases rooted in culture, norms, and values.</li><li>Developing model editing techniques to selectively unlearn unfair instances for bias mitigation in LLMs.</li><li>Designing causal attribution &amp; reasoning bias assessment in closed-format &amp; open-ended LLM outputs.</li><li>Investigating the alignment between cognitive theories &amp; LLM behavior towards aiding bias mitigation.</li></ul>	<b>Fairfax, VA</b>
<b>Sep 2024</b> <b>May 2024</b>	<b>University of Washington   Tech Policy Lab, The Information School</b> <i>Visiting Researcher</i>   Advisor: <a href="#">Aylin Caliskan</a> <ul style="list-style-type: none"><li>Explored factuality, perceptions, stereotyping, &amp; biased decision-making in Visual Question Answering.</li><li>Proposed a framework to identify hidden biased associations in T2T, T2I, &amp; I2T interactions in VLMs.</li><li>Developed a cognition-grounded debiasing approach using in-context learning &amp; instruction-tuning.</li></ul>	<b>Seattle, WA</b>
<b>Aug 2023</b> <b>May 2023</b>	<b>George Mason University   School of Computer Science</b> <i>Graduate Research Assistant</i>   Advisor: <a href="#">Ziwei Zhu</a> <ul style="list-style-type: none"><li>Investigated fairness in transformer-based fake news detection models, applied SHAP and LIME for interpretability, and improved robustness through adversarial attacks and salience-guided input perturbations.</li></ul>	<b>Fairfax, VA</b>
<b>Jul 2022</b> <b>Jun 2021</b>	<b>University of Technology Sydney   School of Computer Science</b> <i>Visiting Scholar</i>   Advisor: <a href="#">Mukesh Prasad</a> <ul style="list-style-type: none"><li>Analyzed disaster-fundraising using event extraction &amp; temporal analysis for the Australian Red Cross.</li><li>Explored AI integration in Aboriginal communities by analyzing language use and cultural contexts.</li><li>Developed transformer-based models to detect web information pollution, cyberbullying &amp; hate speech.</li><li>Applied machine learning to support clinical diagnosis of vertigo by modeling symptom patterns.</li></ul>	<b>Remote</b>

### Selected Publications

[ordered by impact, 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'24]
- [C.4] **Global Voices, Local Biases: Socio-cultural Prejudices across Languages** [PDF | Code]  
[Chahat Raj](#)\*, Anjishnu Mukherjee\*, Ziwei Zhu, Antonios Anastasopoulos  
*Conference on Empirical Methods in Natural Language Processing* [EMNLP'23]
- [C.3] **Breaking Bias, Building Bridges: Evaluation and Mitigation of Social Biases in LLMs via Contact Hypothesis**  
[Chahat Raj](#), Anjishnu Mukherjee, Aylin Caliskan, Antonios Anastasopoulos, Ziwei Zhu [PDF | Code]  
*AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society* [AIES'24]

- [C.2] **SALSA: Saliency-Based Switching Attack for Adversarial Perturbations in Fake News Detection Models** [Chahat Raj](#)\*, Anjishnu Mukherjee\*, Hemant Purohit, Antonios Anastasopoulos, Ziwei Zhu [PDF | Code] [ECIR'24]  
European Conference on Information Retrieval
- [C.1] **True and Fair: Robust and Unbiased Fake News Detection via Interpretable Machine Learning** [PDF | Code] [Chahat Raj](#), Anjishnu Mukherjee, Ziwei Zhu [AIES'23]  
AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society
- [S.4] **VIGNETTE: Socially Grounded Bias Evaluation for Vision-Language Models** [PDF | Code] [Chahat Raj](#), Bowen Wei, Aylin Caliskan, Antonios Anastasopoulos, Ziwei Zhu [in submission]
- [S.3] **Talent or Luck? Evaluating Attribution Bias in Large Language Models** [PDF | Code] [Chahat Raj](#), Mahika Banerjee, Jinhao Pan, Aylin Caliskan, Antonios Anastasopoulos, Ziwei Zhu [in submission]
- [S.2] **Beneath the Surface: How Large Language Models Reflect Hidden Bias** [PDF | Code] Jinhao Pan, [Chahat Raj](#), Ziyu Yao, Ziwei Zhu [in submission]
- [S.1] **Measuring South Asian Biases in Large Language Models** [PDF | Code] Mamnuya Rinki, [Chahat Raj](#), Anjishnu Mukherjee, Ziwei Zhu [in submission]

## Skills

Transformers, Adapter tuning, Instruction tuning, Pretraining, Prompt engineering, In-context learning, Direct Preference Optimization, RLHF, Knowledge distillation, Model compression, LLM Quantization & Optimization, Machine translation, Question answering, Reasoning, Conversational NLP, Summarization, LLM evaluation, Explainability, Adversarial robustness, Safety alignment, Diffusion models, Vision-Language modeling, Multi-agent LLM simulations.

## Selected Projects

- > **AI, Power & Society** Auditing Diffusion Models for Generational and Representational Errors [Report]
- > **Computer Vision** Cultural Image Translation for Human Figures [Report]
- > **AI: Ethics, Policy & Society** Implicit Association Test in Multimodal Generative Models [Report]

## Teaching

- > **CS 108 Introduction to Computer Programming, GMU** Teaching Assistant Spring'24
- > **CS 478 Natural Language Processing, GMU** Teaching Assistant Fall'23
- > **COMP 502 Mathematical Foundations of Computing, GMU** Teaching Assistant Fall'23
- > **CS 112 Introduction to Python Programming, GMU** Teaching Assistant Fall'22, Spring'23

## Talks & Presentations

- > **LLM Ethics - Bias and Mitigation** Invited Lecture - Advanced NLP Oct'24 (GMU, VA)
- > **Breaking Bias, Building Bridges** Talk at AAAI/ACM AIES Oct'24 (San Jose, CA)
- > **A Psychological View to Social Biases in LLMs** Talk at SouthNLP Apr'24 (Emory, GA)

## Student Mentorship

sub-mentoring with Ziwei, Antonis, Aylin

- > **Mamnuya Rinki [MS Thesis, GMU]** Mitigating South Asian Biases in multilingual open-ended LLM generations
- > **Mahika Banerjee [TJ High School]** Measuring reasoning and attribution biases across genders & nationalities
- > **Diwita Banerjee [MS, GMU]** Assessing cuisine biases; scientific figure explanation issues in VLMs
- > **Rinki, Sharanya, Aksh [Advanced NLP]** Intersectional biases in LLMs through Indo-Aryan languages

## 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 and Montreal.
- > **CAHMP Fellowship @ GMU, 2024** Summer research on social biases in question answering.
- > **Best Paper Award @ MASC-SLL 2024** A Psychological View to Social Bias in LLMs.
- > **Summer Graduate Research Award @ GMU, 2023** Explainability and Robustness of Debiasing.
- > **Research Excellence Award @ DTU 2023 & 2022** Received three Commendable Research Awards with INR 150k.
- > **Graduate Scholarship @ DTU 2019** Received INR 297k by AICTE for qualifying GATE.

## Program Committee

- > **Reviewer** AAAI'25, ARR'23-25, COLM'25, TrustNLP'24-25, LTEDI @ EACL'24, CSCW'24, NeurIPS'23, ACM TIST
- > **Sub-Reviewer** NAACL'25, ACL'24, EMNLP'24, AIES'24