Chahat Raj

PhD in Computer Science | George Mason University

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

• NLP • LLMs & VLMs • Multimodality • Multilinguality • Alignment • Responsible AI • Cognitive Science • Ethics • Society • Culture

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, AAAI and ECIR.

Education

Present | **George Mason University** — **Fairfax, Virginia**
Aug 2022 | Ph.D. in Computer Science *(GPA: 3.83)*
Advisors: Ziwei Zhu, Antonios Anastasopoulos

Dec 2025 | **George Mason University** — **Fairfax, Virginia**
Aug 2022 | M.S. in Computer Science

Aug 2021 | **Delhi Technological University** — **Delhi, India**
Aug 2019 | Masters in Information Systems (Thesis)
Advisor: Priyanka Meel

May 2019 | **Indira Gandhi Delhi Technical University for Women** — **Delhi, India**
Aug 2015 | Bachelors in Computer Science & Engineering

Experience

Aug 2025 | **George Mason University, University of Washington** — **Fairfax, VA**
May 2024 | *Graduate Research Assistant | Advisors: Ziwei Zhu, Antonios Anastasopoulos, Aylin Caliskan*
› Societal simulations, model unlearning, and cognitive alignment for bias mitigation.

Sep 2024 | **University of Washington | Tech Policy Lab, The Information School** — **Seattle, WA**
May 2024 | *Visiting Scholar | Advisor: Aylin Caliskan*
› Bias, cognition, and mitigation in Vision Language Models.

Aug 2023 | **George Mason University | School of Computer Science** — **Fairfax, VA**
May 2023 | *Graduate Research Assistant | Advisor: Ziwei Zhu*
› Fairness, interpretability, and robustness in transformer-based fake news detection.

Jul 2022 | **University of Technology Sydney | School of Computer Science** — **Remote**
Jun 2021 | *Visiting Scholar | Advisor: Mukesh Prasad*
› Cyberbullying, AI for Aboriginals, modeling disaster fundraising, and ML for clinical diagnosis.

Selected Publications

full list here

[C.8] **Discovering Bias Associations through Open-Ended LLM Generations** [PDF | Code]
Jinhao Pan, Chahat Raj, Ziwei Zhu
In Proceedings of the AAAI Conference on Artificial Intelligence [**AAAI'26**]

[C.7] **What's Not Said Still Hurts: A Description-Based Evaluation Framework for Measuring Social Bias in LLMs**
Jinhao Pan, Chahat Raj, Ziyu Yao, Ziwei Zhu [PDF | Code]
In Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China [**EMNLP'25**]

[C.6] **Toward Inclusive Language Models: Sparsity-Driven Calibration for Systematic and Interpretable Mitigation of Social Biases in LLMs** [PDF | Code]
Prommy Sultana Hossain, Chahat Raj, Ziwei Zhu, Jessica Lin, Emanuela Marasco
In Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China [**EMNLP'25**]

[C.5] **BiasDora: Exploring Hidden Biased Associations in Vision-Language Models** [PDF | Code]
Chahat Raj, Anjishnu Mukherjee, Aylin Caliskan, Antonios Anastasopoulos, Ziwei Zhu
In Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA [**EMNLP'24**]

[C.4] **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]
Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society [**AIES'24**]

[C.3] **Global Voices, Local Biases: Socio-cultural Prejudices across Languages** [PDF | Code]
Chahat Raj*, Anjishnu Mukherjee*, Ziwei Zhu, Antonios Anastasopoulos
In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore [**EMNLP'23**]

[C.2] **SALSA: Salience-Based Switching Attack for Adversarial Perturbations in Fake News Detection Models**
Chahat Raj*, Anjishnu Mukherjee*, Hemant Purohit, Antonios Anastasopoulos, Ziwei Zhu [PDF | Code]
European Conference on Information Retrieval [**ECIR'24**]

[C.1] **True and Fair: Robust and Unbiased Fake News Detection via Interpretable Machine Learning** [PDF | Code]
Chahat Raj, Anjishnu Mukherjee, Ziwei Zhu
Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society [**AIES'23**]

[S.4] **VIGNETTE: Socially Grounded Bias Evaluation for Vision-Language Models** [PDF | Code]
Chahat Raj, Bowen Wei, Aylin Caliskan, Antonios Anastasopoulos, Ziwei Zhu [**ArXiv**]

[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 [**ArXiv**]

[S.2] **Inoculation Prompting: Implicit Bias Mitigation in Large Language Models** [PDF | Code]
Yongxu Sun, Chahat Raj, Aylin Caliskan [**ArXiv**]

[S.1] **Measuring South Asian Biases in Large Language Models** [PDF]
Mamnuya Rinki, Chahat Raj, Anjishnu Mukherjee, Ziwei Zhu [**ArXiv**]

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 584 Theory/Application Data Mining, GMU** — *Teaching Assistant* — Fall'25
- **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

- **Visual QA & Attribution Biases** — *Posters at MASC-SLL* — Apr'25 (Penn State, PA)
- **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
- **Angela Liu [TJ High School]** — Assessing moral attributional biases in ambiguous contexts in LLMs
- **Daksha Rajagopal [TJ High School]** — Modeling normative ethics and cross-cultural moral reasoning in VLMs
- **Yongxu Sun [MS, UW]** — Inoculation Prompting by in-context learning to debias LLMs' decisions
- **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 across 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** — ARR'23-25, CHI'26, AAAI'26, IASEAI'26, COLM'25, TrustNLP'24-25, LTEDI @ EACL'24, CSCW'24, NeurIPS'23, ACM TIST