Chahat Raj

PhD in Computer Science | George Mason University

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

Present
Aug 2022 | **George Mason University** — **Virginia, USA**
Ph.D. in Computer Science *(GPA: 3.81)*
Advisors: Ziwei Zhu, Antonios Anastasopoulos

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

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

Experience

Present
May 2024 | **University of Washington, George Mason University** — **Fairfax, VA**
Graduate Research Assistant | Advisors: Ziwei Zhu, Antonios Anastasopoulos, Aylin Caliskan
- Simulating societal dynamics using LLMs to audit harms and biases rooted in culture, norms, and values.
- Developing model editing techniques to selectively unlearn unfair instances for bias mitigation in LLMs.
- Designing causal attribution & reasoning bias assessment in closed-format & open-ended LLM outputs.
- Investigating the alignment between cognitive theories & LLM behavior towards aiding bias mitigation.

Sep 2024
May 2024 | **University of Washington | Tech Policy Lab, The Information School** — **Seattle, WA**
Visiting Researcher | Advisor: Aylin Caliskan
- Explored factuality, perceptions, stereotyping, & biased decision-making in Visual Question Answering.
- Proposed a framework to identify hidden biased associations in T2T, T2I, & I2T interactions in VLMs.
- Developed a cognition-grounded debiasing approach using in-context learning & instruction-tuning.

Aug 2023
May 2023 | **George Mason University | School of Computer Science** — **Fairfax, VA**
Graduate Research Assistant | Advisor: Ziwei Zhu
- 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.

Jul 2022
Jun 2021 | **University of Technology Sydney | School of Computer Science** — **Remote**
Visiting Scholar | Advisor: Mukesh Prasad
- Analyzed disaster-fundraising using event extraction & temporal analysis for the Australian Red Cross.
- Explored AI integration in Aboriginal communities by analyzing language use and cultural contexts.
- Developed transformer-based models to detect web information pollution, cyberbullying & hate speech.
- Applied machine learning to support clinical diagnosis of vertigo by modeling symptom patterns.

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: 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
AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society **[AIES'23]**

[S.1] **Beneath the Surface: How Large Language Models Reflect Hidden Bias** [PDF | Code]
Jinhao Pan, Chahat Raj, Ziyu Yao, Ziwei Zhu
Arxiv **[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

› **Question Answering & 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

› **Yongxu Sun [MS, UW]**	Inoculation Prompting by in-context learning to debias LLMs' decisions
› **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**	ACL'25, COLM'25, TrustNLP'25, LTEDI @ EACL'24, TrustNLP'24, CSCW'24, NeurIPS'23, ACM TIST
› **Sub-Reviewer**	NAACL'25, ACL'24, EMNLP'24, AIES'24