Chahat Raj

PhD in Computer Science | George Mason University

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

Research Interests

I work on NLP, and Generative AI, specializing in Large Vision-Language Models focusing on their evaluation, 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 | **George Mason University** | **Fairfax, USA**
Aug 2022 | Ph.D. in Computer Science
Advisors: Ziwei Zhu, Antonios Anastasopoulos

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

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

Experience

Present | **George Mason University | School of Computer Science** | **Fairfax, VA**
May 2024 | *Graduate Research Assistant | Advisors: Ziwei Zhu, Antonios Anastasopoulos*
Projects: Exploring model editing and machine unlearning for bias mitigation in LLMs and VLMs, Developing evaluation frameworks to identify and measure stereotypical associations in multimodal generative tasks. Exploring the alignment between socio-psychological theories and LLM behavior.

Sep 2024 | **University of Washington | The Information School** | **Seattle, WA**
May 2024 | *Visiting Researcher | Advisor: Aylin Caliskan*
Projects: Investigating factuality, hallucination, biased assumptions, and unfair decision-making through Visual Question Answering (VQA) in Generative AI models, Evaluating fairness in text-image interactions, Bias mitigation in LLMs using In-context Learning and Instruction Tuning.

Aug 2023 | **George Mason University | School of Computer Science** | **Fairfax, VA**
May 2023 | *Graduate Research Assistant | Advisor: Ziwei Zhu*
Projects: Bias and Fairness in transformer models for fake news detection, model interpretability with SHAP and LIME, enhancing robustness through adversarial attacks and salience-based perturbations.

Jul 2022 | **University of Technology Sydney | School of Computer Science** | **Remote**
Jun 2021 | *Visiting Scholar | Advisor: Mukesh Prasad*
Projects: 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.

Aug 2021 | **Indian Institute of Management Raipur** | **Remote**
May 2021 | *Researcher | Advisor: Manojit Chattopadhyay*
Projects: Cryptocurrencies' price prediction using variational autoencoders and false nearest neighbor approximation, Analysing causal relationships between stock market behavior and societal events.

Aug 2022 | **Delhi Technological University** | **Delhi, India**
Aug 2019 | *Graduate Student Researcher | Advisor: Priyanka Meel*
Projects: Developing deep learning frameworks to detect antisocial web content (fake news, infodemic, misinformation, disinformation, rumor, death hoax, and clickbait), Infodemic impact modeling.

Skills

NLP, Generative AI, Multimodal Large Language Models, Model Training, Evaluation & Validation, Instruction-Tuning, Prompt Engineering, LLM-Human Alignment, Vision, Ethics and Bias Mitigation, Interpretability, Efficient Fine-Tuning (LoRA, PEFT, QLoRA), Hallucination Detection & Mitigation, Diffusion Models, Contrastive Learning, LLM Quantization & Optimization (bitsandbytes), Self-Correction & Reflexion in LLMs, Multi-Step Reasoning & Planning (Chain-of-Thought).

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: Salience-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]
Chahat Raj*, Anjishnu Mukherjee*, 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]**

Talks

"LLM Ethics - Bias and Mitigation"

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

"Breaking Bias, Building Bridges"

› Talk at AAAI/ACM AIES — October 2024 (San Jose, 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 $(\times 2)$ **& 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
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

Academic Service

Reviewer TrustNLP'25, LTEDI @ EACL'24, TrustNLP'24, CSCW'24, NeurIPS'23, ACM Transactions on Intelligent Systems and Technology'24 & '23, Elsevier & Springer Journal papers

Sub-Reviewer NAACL'25, ACL'24, EMNLP'24, AIES'24

Student Mentorship

› Mamnuya Rinki [MS, GMU]
› Mahika Banerjee (Sub-mentoring with Dr. Ziwei Zhu) [TJ High School]