Arka Dutta

Rochester, NY| ad2688@rit.edu| 585-505-9739| Google Scholar |GitHub | Website| LinkedIn

EDUCATION

Rochester Institute of Technology — Rochester, NY
Doctor of Philosophy (Ph.D.) — Aug 2023 - Present
Computing and Information Sciences

Kalyani Government Engineering College — Kalyani, WB
Bachelor of Technology (B. Tech) — Aug 2019 - Jul 2023
Computer Science and Engineering

WORK EXPERIENCE

Intuit — Mountain View, CA
AI Science Intern — May 2025 - Aug. 2025
- Supervised by Na Xu

Rochester Institute of Technology — Rochester, NY
Graduate Research Assistant — Aug. 2023 – Present
- PhD student advised by Prof. Dr. Ashiqur R. KhudaBukhsh. Research focus on Responsible AI and equitable AI systems.
- Published papers in top conferences including papers in IJCAI 2024 (Acceptance rate~ 20%), AAAI 2025(Acceptance rate~ 20%); research covered in widely circulated AI Magazines like Montreal AI Ethics with an invited talk from 5+ premier research institutes.

SKILLS

- Machine Learning (ML), Deep Learning, Artificial Intelligence (AI), Natural Language Processing (NLP), Large Language Model (LLM), Explainable AI (XAI), Reinforcement Learning with Human Feedback (RLHF), Natural Language Understanding, Generative AI, Responsible AI, Optimization, Statistical Inference, Predictive Modeling, Data Visualization, Multimodal AI,
- Python, C/C++, MATLAB, SQL, Transformers, PyTorch, TensorFlow, NumPy, Pandas, NLTK, Spacy, HuggingFace, SHAP

ACHIEVEMENTS

Language Science Student Excellence Award: One of the two recipients of the award among 1000+ students university-wide in 2024; for the research in Responsible AI and NLP.
Common Admission Test (CAT, 2022): Ranked 98.97 %ile among 100k+ students in highly competitive CAT exam held in India as premier B-School admission test. Offered an admission from Indian Institute of Management, Indore (IIMI) PGP program; one of the most prestigious B-Schools in the world.
Regional Mathematics Olympiad (2018): Qualified for prestigious RMO, WB as one of the top 100 in state; demonstrating advanced level aptitude in Discrete Mathematics, Combinatorics, Number Theory, Algebra, and Geometry.
Jagadish Bose National Science Talent Scholarship (2017): Awarded prestigious JBNSTS scholarship for excellence in science scored in the top 500 among 10000+ test-takers.
Media Attention: (1) Research featured in WIRED, a widely circulated media outlet: Link
(2) Research featured in Montreal AI Ethics Newsletter, a widely circulated AI Ethics newsletter: Link

RESEARCH

PUBLICATIONS

- **A. Dutta∗,** S. M. Sualeh Ali∗, U. Naseem, and A. R. KhudaBukhsh: ***Towards a Bipartisan Understanding of Peace and Vicarious Interactions.*** in Proceedings of International Joint Conference on Artificial Intelligence, AI for Good (IJCAI-25).
- **A. Dutta,** A. Priyanshu, and A. R. KhudaBukhsh: ***All You Need Is S P A C E: When Jailbreaking Meets Bias Audit and Reveals What Lies Beneath the Guardrails (Student Abstract),*** in Proceedings of AAAI Conference on Artificial Intelligence (AAAI-25*),* Pages: 29353-29355, Oral Presentation (11.6%) [PDF]
- **A. Dutta∗,** A. Khorramrouz∗, S. Dutta, and A. R. KhudaBukhsh: ***Down the Toxicity Rabbit Hole: A Novel Framework To Bias Audit Large Language Models.*** in Proceedings of International Joint Conference on Artificial Intelligence, AI for Good (IJCAI-24). Pages 7242-7250, Oral Presentation (15%) [PDF]

- **A. Dutta,** A. Baral, S. Kundu, S. Biswas, K. Dasgupta, and Hasanujjaman: ***Classification of Cricket Shots from Cricket Videos Using Self-Attention Infused CNN-RNN (SAICNN-RNN).*** In Proceedings of CICBA 2023. Springer Link. DOI: 10.1007/978-3-031-48876-4_24

MANUSCRIPTS

- R. Magu, **A. Dutta**, S. Kim, A.R. KhudaBukhsh, and M. De Choudhury: ***Navigating the Rabbit Hole: Emergent Biases in LLM-Generated Attack Narratives Targeting Mental Health Groups,*** (In Review - arXiv version)[PDF]
- **A. Dutta**, U. Jaimini, U. Bhatt, S S Muthuselvam, A. Das, and A. R. KhudaBukhsh: ***A Large Scale Social Web Audit of AI Generated Text Detection Systems.*** (In Review - arXiv version)
- **A. Dutta,** R. Fayyazi, S. Yang, and A. R. KhudaBukhsh: ***How Can You Tell if Your Large Language Model Could Be a Closet Antisemite? A Framework to Bias Audit Large Language Models.*** (In Review - arXiv version)
- A. R. KhudaBukhsh, **A. Dutta**, and A. Mukherjee: ***Counterbalancing Hate with Positivity: A Survey of Counterspeech.*** (In Review - arXiv version)