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

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 one paper in IJCAI 2024 (Acceptance rate~ 20%); research covered in widely circulated AI Magazines like Montreal AI Ethics with an invited talk from 5+ premier research institutes.

Indian Statistical Institute — Kolkata, India
NLP Research Intern — Jul. 2022 – Oct. 2022

- Supervised by Prof. Dr. Utpal Garain: Devised an approach to extract hate-phrases for hate-speech classification in Bengali and Hindi, two low-resource languages using HASOC dataset.
- Used Language Models like RoBERTa, AlBERT and XAI methods like Deletion-intervention, game theory-based Feature Ablation to extract hate speech that resulted in identifying ~73% hate-phrases towards explaining the final classifier prediction. This resulted in a ~12% improvement in classification accuracy in hate-speech classification than the previous baseline.

Jadavpur University — Remote
Research Intern — Dec. 2021 – Jul 2022

- Under the supervision of Dr. Arnab Raha, Staff Research Scientist, Intel AI: Systemized a Distributed Deep Neural Network approach with Adaptive Workload Partitioning along low-computing edge devices to minimize the latency due to cloud transfer for faster and resource-friendly computation and communication.
- Used Neural Network based optimization to achieve a 2.3x dip in latency improving computation efficiency, which could translate into significant cost savings for edge computing applications.

SKILLS

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

ACHIEVEMENTS

Language Science Student Excellence Award: One of the only two recipients of the award among 1000+ students university-wide in 2024; for the research in bias auditing Large Language Models 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 premier 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, Geometry, and Optimization.
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: Research featured in Montreal AI Ethics Newsletter, a widely circulated AI Ethics newsletter.

RESEARCH

PUBLICATIONS

- **A. Dutta,** A. Priyanshu, and A. R. KhudaBukhsh: What Lies Beneath the Guardrails? Jailbreaking Meeting Bias Audit, Student Abstract, AAAI Conference on Artificial Intelligence (AAAI-25*),* Oral Presentation (11.6%)
- **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 the Thirty-Third International Joint Conference on Artificial Intelligence, AI for Good. Pages 7242-7250. URL: https://doi.org/10.24963/ijcai.2024/801 , Poster (15%)
- **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

- **A. Dutta∗,** S. M. Sualeh Ali∗, U. Naseem, and A. R. KhudaBukhsh: Towards a Bipartisan Understanding of Peace and Vicarious Interactions. in review
- **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. arXiv preprint, in review
- A. R. KhudaBukhsh, **A. Dutta**, and A. Mukherjee: Counterbalancing Hate with Positivity: A Survey of Counterspeech. arXiv preprint