Salman Rahman

salmanrahman.net

Education

2023–Present Ph.D. in Computer Science (AI/NLP), *GPA: 4.00/4.00*, University of California, Los Angeles
— Advisor: Saadia Gabriel

2021–2022 M.S. in Applied Statistics and Data Science, *GPA: 4.00/4.00*, University of Texas RGV

Selected Research Projects

Project 1 **Scalable Red Teaming Framework for Multi-Turn Jailbreaking**
— Multi-agent red-teaming framework designed to systematically generate multi-turn jailbreak attacks and provide large-scale open-source safety resources for robust, real-time safeguarding of interactive LLM systems[pdf summary]

Project 2 **Emergent Social Behavior in Multiagent LLMs**
— Developing a multi-agent simulation platform to explore emergent social behaviors and risks in LLM-based agents through realistic network interactions and game-theoretic modeling[pdf summary]

Project 3 **Scalable Oversight & Generalization**
— Developing human supervision methods for advanced AI systems via multi-agent LLM debate and consultation frameworks, focusing on complex tasks where ground truth is difficult to verify [debate app]

— Systematic evaluation and targeted fine-tuning of clinical language models to enhance generalization across diverse healthcare settings[arXiv]
— Analysis of how data factors (such as sample size imbalance, covariate shift, concept shift, and omitted variables) and model complexity contribute to disparities in LIME explanations[FAccT]
— Comprehensive assessment of vision model robustness across convolution, attention, hybrid, sequence-based, and network-based architectures under out-of-distribution settings[arXiv]

Project 4 **AI for Social Good**
— Large-scale Google Street View image analysis demonstrates that neglecting domain knowledge and mediators biases built-environment interventions for obesity and diabetes, highlighting the importance of causal frameworks in AI-driven health decisions[PNAS]
— Ensemble machine learning methods for landslide prediction, optimizing spatial agreement and reducing model uncertainties[remote sensing]
— Machine learning and optimization approaches for sustainable resource management in Bangladesh, focusing on waste-to-energy[journal], solar power[journal], and agricultural systems[journal]

Industry Experience

2024 **Multimodal AI Research Intern, Apple**
— Developed a pipeline to generate high-quality, task-specific synthetic data for fine-tuning Apple's MM1 multimodal model on specialized computer vision tasks

Work Experience

2024–Present Research Fellow, UCLA Computer Science

2022–2023 Research Assistant, NYU Computer Science

2022 Teaching Assistant, University of Texas, Mathematics & Statistics
— STAT 3337: Probability and Statistics; STAT 3301: Applied Statistics

2021 Research Assistant, University of Texas, Mathematics & Statistics

Selected Publications

2024 Understanding Disparities in Post Hoc Machine Learning Explanations
— Vishwali Mhasawade, **Salman Rahman**, Zoe Haskell-Craig, Rumi Chunara
ACM Conference on Fairness, Accountability, and Transparency (FAccT)

2024 Generalization in Healthcare AI: Evaluation of a Clinical Large Language Model
— **Salman Rahman**, Lavender Yao Jiang, Saadia Gabriel, Yindalon Aphinyanaphongs, Eric Karl Oermann, Rumi Chunara
arXiv Preprint

2024 Utilizing Big Data Without Domain Knowledge Impacts Public Health Decision-Making
— Miao Zhang, **Salman Rahman**, Vishwali Mhasawade, Rumi Chunara
Proceedings of the National Academy of Sciences (PNAS)

2020 Improving Spatial Agreement in Machine Learning-Based Landslide Susceptibility Mapping
— Mohammed Sarfaraz Gani Adnan, **Salman Rahman**, Nahian Ahmed, Bayes Ahmed, Md. Fazleh Rabbi, Rashedur M. Rahman
Remote Sensing

Awards

2024 PhD Fellowship, UCLA

2021 Presidential Graduate Research Scholarship, University of Texas

2021 LaunchPad Ideas Competition Grand Prize, Blackstone