Jwala Dhamala

LinkedIn: LinkedIn
Webpage: jwaladhamala.com
jwala.dhamala@gmail.com
585.314.9794

Research interests

Large language models, responsible AI, model safety, natural language understanding, model evaluation.

Education

Ph.D. in Computing and Information Sciences 2014 - 2020
Rochester Institute of Technology, Rochester, NY, US GPA: 3.93/4.00
Advisor: Dr. Linwei Wang

B.E. in Computer Engineering 2008 - 2012
Pulchowk Campus, Tribhuvan University, Nepal with Distinction

Experience

Senior Applied Scientist 2022 - Present
AGI, Amazon
Research focus: Evaluation and benchmarking RAI in large multimodal language models, including benchmark creation, metric design, and model evaluation. Human and automated red-teaming. Jailbreaking attack. Model safety and robustness for low-resource multilingual languages. Data safety, e.g., PII removal and toxic content removal.

Applied Scientist 2021 - 2022
Alexa AI - Natural Language Understanding, Amazon
Research interest: Responsible AI for NLU models. Discovering underperformance cohorts and mitigating performance gaps through model training and evaluations.

Research Scientist 2019 - 2021
Alexa AI - Natural Language Understanding, Amazon

Research Assistant 2014 - 2019
Computational Biomedicine Lab
Rochester Institute of Technology, NY, US
Research focus: Machine/deep learning approaches to integrate measurements with physics-based simulations for probabilistic personalization of the simulation models. Experience with machine learning methods like Gaussian processes, Bayesian optimization and MCMC; and deep learning methods like variational auto-encoders (VAE) and geometric deep learning.

Research Intern 2018
Philips Healthcare, Cambridge, MA, US
Research focus: Unsupervised representation learning and similarity assessment of multi-variate time-series physiological signals. Experience with RNNs, LSTMs and approximate nearest neighbor methods.

Software Engineer 2012 – 2014
Business Intelligence Department
Logic Information Systems, Nepal
Focus: Worked and lead projects on ETL for data warehousing and statistical data analysis for business intelligence dashboards. Designed and conducted training sessions for interns.

Select Conference articles

Tree-of-Traversals: A Zero-Shot Reasoning Algorithm for Augmenting Black-box Language Models with Knowledge Graphs
E. Markowitz, A. Ramakrishna, **J. Dhamala**, N. Mehrabi, C. Peris, R. Gupta, K. Chang, A. Galstyan
Association for Computational Linguistics (ACL), 2024

Tokenization Matters: Navigating Data-Scarce Tokenization for Gender Inclusive

Language Technologies
A. Ovalle, N. Mehrabi, P. Goyal, **J. Dhamala**, K. Chang, A. Galstyan, R. Zemel, Y. Pinter, R. Gupta
NAACL Findings 2024

MICo: Preventative Detoxification of Large Language Models through Inhibition Control
R. Siegelmann, N. Mehrabi, P. Goyal, P. Goyal, L. Bauer, **J. Dhamala**, A. Galstyan, R. Gupta, R. Ghanadan
NAACL Findings 2024

Resolving Ambiguities in Text-to-Image Generative Models
N. Mehrabi, P. Goyal, A. Verma, **J. Dhamala**, V. Kumar, Q. hu, K. Chang, R. Zemel, A. Galstyan, R. Gupta
Association for Computational Linguistics (ACL), 2023

Im fully who I am: Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language Generation
A. Ovalle, P. Goyal, **J. Dhamala**, Z. Jaggers, K. Chang, A. Galstyan, R. Zemel, R. Gupta
FaccT 2023

On the Intrinsic and Extrinsic Fairness Evaluation Metrics for Contextualized Language Representations
Y. Trista Cao, Y. Pruksachatkun, K. Chang, R. Gupta, V. Kumar, **J. Dhamala**, A. Galstyan
Association for Computational Linguistics (ACL), 2022

Mitigating Gender Bias in Distilled Language Models via Counterfactual Role Reversal
U. Gupta, **J. Dhamala**, V. Kumar, A. Verma, Y. Pruksachatkun, S. Krishna, R. Gupta, K. Chang, G. Steeg & A. Galstyan
Association for Computational Linguistics (ACL findings), 2022

Measuring Fairness of Text Classifiers via Prediction Sensitivity
S. Krishna, R. Gupta, A. Verma, **J. Dhamala**, Y. Pruksachatkun & K. Chang
Association for Computational Linguistics (ACL), 2022

Does Robustness Improve Fairness? Approaching Fairness with Word Substitution Robustness Methods for Text Classification
Y. Pruksachatkun, S. Krishna, **J. Dhamala**, R. Gupta & K. Chang
North American Chapter of the Association for Computational Linguistics (NAACL findings), 2021

BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation
J. Dhamala*, T. Sun*, V. Kumar, S. Krishna, Y. Pruksachatkun, K. Chang & R. Gupta
ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT), 2021

Learning Geometry-Dependent and Physics-Based Inverse Image Reconstruction
X. Jiang, S. Ghimire, **J. Dhamala**, Z. Li, P. K. Gyawali & L. Wang
Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2020

Bayesian Optimization on Large Graphs via a Graph Convolutional Generative Model: Application in Cardiac Model Personalization
J. Dhamala, S. Ghimire, J. L. Sapp, B. M. Horáček & L. Wang
Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2019
early acceptance (selection rate ∼ 15%), finalist for young scientist award

Improving Generalization of Deep Networks for Inverse Reconstruction of Image Sequences
S. Ghimire, P. K. Gyawali, **J. Dhamala**, J. L. Sapp, J. L., Horáček, M., and Wang, L.
Information Processing in Medical Imaging (IPMI), 2019
oral presentation

High-dimensional Bayesian Optimization of Personalized Cardiac Model Parameters via an Embedded Generative Model
J. Dhamala, S. Ghimire, J. L. Sapp, B. M. Horáček & L. Wang
Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2018
oral presentation, finalist for young scientist award (selection rate $\sim$ 1%)

Generative Modeling and Inverse Imaging of Cardiac Transmembrane Potential
S. Ghimire, **J. Dhamala**, P. K. Gyawali, J. L. Sapp, B. M. Horáček & L. Wang
Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2018

Quantifying the Uncertainty in Model Parameters using Gaussian Process-based Markov Chain Monte Carlo: an Application to Cardiac Electrophysiological Models
J. Dhamala, J. L. Sapp, B. M. Horáček & L. Wang
Information Processing in Medical Imaging (IPMI), 2017
acceptance rate $\sim$ 30%

Overcoming Barriers to Quantification and Comparison of Electrocardiographic Imaging Methods: a Community-based Approach
S. Ghimire, **J. Dhamala**, J. Coll-Font, J. D. Tate, M. S. Guillem, D. H. Brooks, R. S. MacLeod & L. Wang
Computing in Cardiology (CinC), 2017

The Consortium for Electrocardiographic Imaging
J. Coll-Font, **J. Dhamala**, D. Potyagaylo, W. H. Schulze, J. D. Tate, M. S. Guillem, P. Van Dam, O. Dossel, D. H. Brooks & R. S. Macleod
Computing in Cardiology (CinC), 2016

Spatially-adaptive Multi-scale Optimization for Local Parameter Estimation: Application in Cardiac Electrophysiological Models
J. Dhamala, J. L. Sapp, B. M. Horáček & L. Wang
Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2016
early acceptance, selection rate $\sim$ 10%

Journal articles

Fast Posterior Estimation of Cardiac Electrophysiological Model Parameters via Bayesian Active Learning
M. Zaman, **J. Dhamala**, P. Bajracharya, H. J. Arevalo, J. Sapp, M. Horáček, K. C. Wu, N. A. Trayanova & L. Wang
Medical Image Analysis (MedIA), 2020, invited

Embedding High-dimensional Bayesian Optimization via Generative Modeling: Parameter Personalization of Cardiac Electrophysiological Models
J. Dhamala, H. J. Arevalo, J. Sapp, M. Horáček, K. C. Wu, N. A. Trayanova & L. Wang
Medical Image Analysis (MedIA), 2020, invited

Quantifying the Uncertainty in Model Parameters using Gaussian Process-based Markov Chain Monte Carlo in Cardiac Electrophysiology
J. Dhamala, H. J. Arevalo, J. Sapp, M. Horáček, K. C. Wu, N. A. Trayanova & L. Wang
Medical Image Analysis (MedIA), 2018

Multivariate Time-series Similarity Assessment via Unsupervised Representation Learning and Stratified Locality Sensitive Hashing: Application to Early Acute Hypotensive Episode Detection
J. Dhamala, E. Azuh, A. Al-Dujaili, J. Rubin & U. M. O'Reilly
IEEE Sensors Letters, 2018

Spatially Adaptive Multi-scale Optimization for Local Parameter Estimation in Cardiacelectrophysiology
J. Dhamala, H. J. Arevalo, J. Sapp, M. Horáček, K. C. Wu, N. A. Trayanova & L. Wang
IEEE Transactions on Medical Imaging (IEEE TMI), 2017

Patent

Cohort determination in natural language processing

R. Gupta, **J. Dhamala**, A. Verma, Q. Ye, M. Dabhi, S. Veeravanallur, S. Matsoukas, M. Gens, S. Razavi, A. Khatri, P. Natarajan
US Patent 2024

Model Configuration
R. Gupta, **J. Dhamala**, M. Gens, S. Midha, J. Yuen, D. Ibtesham, W. Hamza, X. Zhang, M. Arafat
US Patent 2024

Technical skills

Languages: Python, MATLAB
Deep Learning Framework: PyTorch
Misc: Bash scripting, LaTeX typesetting, Git
Basic familiarity: R, Java, C, C++, HTML, PHP, MySQL

Workshop articles

Evaluating the Effectiveness of Efficient Neural Architecture Search for Sentence-Pair Tasks
A. MacLaughlin, **J. Dhamala**, A. Kumar, S. Venkatapathy, R. Venkatesan & R. Gupta
Workshop on Insights from Negative Results in NLP, EMNLP, 2021

High-dimensional Bayesian Optimization of Personalized Cardiac Model Parameters via an Embedded Generative Model
J. Dhamala, S. Ghimire, J. L. Sapp, B. M. Horáček & L. Wang
Women in Machine Learning (WiML), 2018

Multivariate Time-series Similarity Assessment via Unsupervised Representation Learning and Stratified Locality Sensitive Hashing: Application to Early Acute Hypotensive Episode Detection
J. Dhamala, E. Azuh, A. Al-Dujaili, J. Rubin, and U. M. O'Reilly.
NeurIPS Machine Learning in Healthcare (NeurIPS ML4H), 2018

Scholarships & awards

Travel Grant, NeurIPS Machine learning for Health Workshop (ML4H)	2018
Travel Grant, Woman in Machine Learning (WiML)	2018
Travel Grant, MICCAI	2016, 2018
IPMI Scholarship for Junior Scientists, IPMI	2017
GCCIS Student Grant, Rochester Institute of Technology	2017
Graduate Student Travel Award, Rochester Institute of Technology	2015
Women in Engineering Scholarship, University Grants Commission, Nepal	2010-2011
The College Fellowship Scholarship, Granted 8/8 semesters based on academic merit, Tribhuvan University	2008-2012
Golden Jubilee Scholarship, Government of India	2008-2012
Full-tuition waiver, Based on the performance on a countrywide university entrance examination, Institute of Engineering, Tribhuvan University	2008-2012
Mahatma Gandhi Scholarship, Government of India	2006-2007

Professional activities

Area Chair

Conference: ARR February 2025	2025

Reviewing

Conference: ACL ARR reviews	2021-present
Conference: NeurIPS	2021
Journal: Data Mining and Knowledge Discovery (Springer)	2021
Journal: Engineering Applications of Artificial Intelligence (Elsevier)	2021
Conference: MICCAI	2017-2021
Workshop: Woman in Machine Learning (WiML)	2018
Journal: IEEE Sensors Letters	2018
Journal: Journal of Biomedical and Health Informatics	2018
Journal: Engineering Applications of Artificial Intelligence	2021

Organization

TrustNLP: Workshop on Trustworthy Natural Language Processing North American Chapter of the Association for Computational Linguistics (NAACL)	2021 - 2025

Workshop on Measures and Best Practices for Responsible AI 2021
ACM SIGKDD Conference on Knowledge Discovery and Data Mining

Pre-orientation program 2017
Woman in Computing, Rochester Institute of Technology

Workshop on Premature Ventricular Contractions Localization 2016, 2017
Computing in Cardiology, Consortium of Electrocardiographic Imaging

LOCUS - Technological Festival 2012
Institute of Engineering, Pulchowk Campus

Invited talks

Fairness in Large-scale Languge Models
Twitch, Responsible AI Tech Talk Series
(Online Event)

Fairness in Open-ended Language Generation
Workshop on Women in Science: Status, Challenges, Opportunities and Innovations, 2021
NEGAAS, Kathmandu, (Online Event)

Applications of Artificial Intelligence for Social Good
Women in Data Science (WiDS), 2021
Kathmandu, Nepal (Online Event)

Applications of Deep Learning to Multi-scale Physics-based Simulators
National Workshop on Machine Learning and Data Science, 2020
Kathmandu, Nepal (Online Event)

Model Personalization and Uncertainty Quantification in Cardiac Electrophysiological Models
Ph.D. Colloquium Series, 2018
College of Computing and Information Sciences, Rochester Institute of Technology
Rochester, NY, US

Personalization and Uncertainty Quantification in Cardiac Electrophysiological Models
Signal Processing Imaging Reasoning and Learning (SPIRAL) Seminar, 2018
Northeastern University, Boston, MA, US