Kshitish Ghate

kshitishghate.github.io
kghate@cs.washington.edu

Paul G. Allen School of Computer Science and Engineering
University of Washington
185 E Stevens Way NE, Seattle, WA 98195

◇ RESEARCH INTERESTS

– Interactive evaluation and improvement of AI agents' social reasoning capabilities.
– Aligning models to diverse principles and values in human-agent/multi-agent interactions.

◇ EDUCATION

2025 – Present **University of Washington**, Seattle, WA
Ph.D. in Computer Science & Engineering
Advisors: Aylin Caliskan, Tadayoshi Kohno
Relevant Coursework: Social Reinforcement Learning, Graduate Security and Privacy Seminar

2023 – 2025 **Carnegie Mellon University**, Pittsburgh, PA
Masters in Language Technologies
Advisor: Mona Diab
Cumulative GPA: 4.0/4.0
Relevant Coursework: Advanced NLP, Introduction to Machine Learning (graduate level), Multimodal ML, On-Device ML, LLM Applications, Question Answering

2018 – 2023 **Birla Institute of Technology and Science, Pilani**, Goa, India
Bachelor of Engineering in Computer Science and Master of Science in Economics
Cumulative GPA: 9.0/10.0
Relevant Coursework: Artificial Intelligence, Data Structures and Algorithms, Database Management Systems, Operating Systems, Foundations of Data Science, Game Theory, Machine Learning, Object Oriented Programming, Probability and Statistics

◇ SKILLS

Languages	Python, R, STATA, C/C++, Matlab, Java, SQL
Toolkits/Cloud	PyTorch, HF Transformers, Pandas, NumPy, vLLM, GCP, AWS, Azure, WandB, Docker, Slurm

◇ PUBLICATIONS

— Preprints & Manuscripts —

(2025) **Kshitish Ghate**, Andy Liu, Devansh Jain, Taylor Sorensen, Atoosa Kasirzadeh, Aylin Caliskan, Mona T. Diab, Maarten Sap. "EVALUESTEER: Measuring Reward Model Steerability Towards Values and Preferences."

(2025) Andy Liu, **Kshitish Ghate**, Mona Diab, Daniel Fried, Atoosa Kasirzadeh, and Max Kleiman-Weiner. "Generative Value Conflicts Reveal LLM Priorities."

— Select Conference and Journal Publications —

2025 **Kshitish Ghate**, Tessa Charlesworth, Mona T. Diab, and Aylin Caliskan. "Biases Propagate in Encoder-based Vision-Language Models: A Systematic Analysis From Intrinsic Measures to Zero-shot Retrieval Outcomes." *Findings of the Association for Computational Linguistics, 2025.*

2025 Jingwen Cheng, **Kshitish Ghate**, Wenyue Hua, William Yang Wang, Hong Shen, and Fei Fang. "REALM: A Dataset of Real-World LLM Use Cases." *Findings of the Association for Computational Linguistics, 2025.*

2025 **Kshitish Ghate***, Isaac Slaughter*, Kyra Wilson, Mona Diab, and Aylin Caliskan. "Intrinsic Bias is Predicted by Pretraining Data and Correlates with Downstream Performance in Vision-Language Encoders." *Nations of the Americas Chapter of the Association for Computational Linguistics, 2025.*

2024 Tessa Charlesworth, **Kshitish Ghate**, Aylin Caliskan, and Mahzarin R. Banaji. "Extracting intersectional stereotypes from embeddings: Developing and validating the Flexible Intersectional Stereotype Extraction procedure." *PNAS Nexus, 2024.*

EXPERIENCE

09/25 – Present **University of Washington**, Seattle, WA
Graduate Research Assistant, Paul G. Allen School of Computer Science and Engineering
Advisors: Aylin Caliskan, Tadayoshi Kohno

– Leading development of ecologically grounded, multi-turn adaptive benchmarks to evaluate and align multi-LLM agents; designing neuro-symbolic and reinforcement learning methods to train LLMs to reason in user-agent and multi-agent settings.

09/23 – 08/25 **Carnegie Mellon University**, Pittsburgh, PA
Graduate Research Assistant, Language Technologies Institute
Advisors: Mona Diab, Maarten Sap

– Developed EVALUESTEER, a controlled synthetic benchmark evaluating LLM and reward-model (RM) steerability to user values and styles, revealing a >25-point steerability gap in RMs.
– Created CONFLICTSCOPE, an fully-automated evaluation pipeline generating 1K+ value-conflict scenarios, and improving alignment consistency under conflict by 14% through prompt steering.
– Investigated the relationship between intrinsic biases in 131 CLIP models, their pretraining factors and demonstrated the propagation of representation biases to downstream retrieval tasks.
– Devised a harm reduction framework for reducing hallucinations and improving accuracy in LLM responses used for clinical decision-making through counterfactual synthetic data generation.
– Introduced a risk taxonomy for Personal Information memorization in LLMs and improved existing detectors with 90% better FPR.

08/22 – 12/22 **Amazon**, Bangalore, India
Applied Scientist Intern, Alexa – AI Natural Language Understanding
Supervisors: Anurag Dwarakanath, Anjali Shenoy

– Implemented a novel training methodology and model architecture, drawing from Curriculum Learning literature, to address problem of classifying long tail data in NLU tasks.
– Achieved 5% improvement in F1 score and Intent Classification accuracy by applying a holistic sample difficulty metric in training.

TEACHING

Fall 2024 Large Language Models: Methods and Applications (11-667), Teaching Assistant, CMU
Spring 2022 Applied Econometrics, Teaching Assistant, BITS Pilani

HONORS & AWARDS

2018 - 2023 National Talent Search Examination (NTSE) Scholarship, NCERT