Gaurav R. Ghosal

305 Stratford Avenue Avenue
Pittsburgh, Pennsylvania, 15213
gghosal@andrew.cmu.edu
gauravrghosal@gmail.com
Github : github.com/GGGG1020
+1-510-778-5797

Education

Carnegie Mellon University
Doctor of Philosophy, Machine Learning *Aug' 23 - Dec' 28 (Expected)*
Advisor: Professor Aditi Raghunathan
Graduate Coursework: Intermediate Statistics (A+), Advanced Machine Learning Theory (A), Advanced Statistical Theory (A), Theoretical and Empirical Foundations of Machine Learning (A)
Research Areas: Foundation Model Reliability, Model Editing and Unlearning, Fact-based Reasoning in Large Language Models.

University of California, Berkeley
Bachelor of Science, EECS *Aug' 19 - Dec' 22*
GPA: 4.0/4 (Overall)
Activities and Societies: Regents' and Chancellor's Scholars Association, Eta Kappa Nu (EECS Honors Society), EECS Honors Program
Graduate Coursework: Deep Reinforcement Learning (A+), Computational Principles for High-dimensional Data Analysis (A+), Convex Optimization and Approximation (A), Theoretical Statistics (A)
Undergraduate Coursework: Machine Learning (A+), Computer Security (A+), Optimization Models in Engineering (A), Algorithms (A), Real Analysis (A+), Differential Topology (A+)

Research Interests

Foundation Model Reliability, Model Editing and Unlearning, Fact-based Reasoning in Large Language Models.

Publications and Pre-prints

Khurram Yamin, Shantanu Gupta, **Gaurav R. Ghosal**, Zachary Lipton, and Bryan Wilder. "Failure Modes of LLMs for Causal Reasoning on Narratives". *Under Review* at International Conference of Machine Learning (2025).

Gaurav R. Ghosal, Tatsunori Hashimoto, and Aditi Raghunathan. "Understanding Finetuning for Factual Knowledge Extraction." International Conference on Machine Learning (ICML), 2024.

Evan Ellis, **Gaurav R. Ghosal**, Stuart Russell, Anca Dragan, and Erdem Biyik. "A Generalized Acquisition Function for Preference-based Reward Learning." International Conference on Robotics and Automation (ICRA), 2024.

Gaurav R. Ghosal, Amrith Setlur, Daniel S. Brown, Anca D. Dragan, and Aditi Raghunathan. "Contextual Reliability: When Different Features Matter in Different Contexts." International Conference on Machine Learning (ICML), 2023.

Gaurav R. Ghosal, Matthew Zurek, Daniel S. Brown, and Anca D. Dragan. "The Effect of Modeling Human Rationality Level on Learning Rewards from Multiple Feedback Types." AAAI Conference on Artificial Intelligence (AAAI), 2023.

Gaurav R. Ghosal, and Reza Abbasi-Asl. "Multi-Modal Prototype Learning for Interpretable Multivariable Time Series Classification." *arXiv preprint arXiv:2106.09636* (2021).

Gaurav R. Ghosal, et al. "A Deep Deterministic Policy Gradient Based Network Scheduler For Deadline-Driven Data Transfers." *2020 IFIP Networking Conference (Networking)*. IEEE, 2020.

Awards & Achievements

Selected for the **EECS Honors Program** for undergraduate EECS and CS students at University of California, Berkeley.
Awarded the **Edward Kraft Award Fall 2019** for first semester freshmen with a 4.0 GPA.
Awarded the **Regents' and Chancellor's Scholarship 2019** for incoming freshmen at University of California, Berkeley.

Research Experience

Reliability in Large Language Models
Carnegie Mellon University Machine Learning Department *Aug '23 - Present*

- Working in Professor Aditi Raghunathan's group as a second year Ph.D. student.
- Published first-author work on investigating the role of fine-tuning in factuality of large language models.
- Current research on improving the localization of memorization and knowledge in language models to enable reliable editing and unlearning.
- Current research on investigating counterfactual reasoning using parametric knowledge in large language models.

Human-in-the-Loop Learning and Robustness
Berkeley Artificial Intelligence Research *Aug '21 - Present*

-Worked in Professor Anca Dragan's Lab under the supervision of post-docs Dr. Daniel Brown and Dr. Aditi Raghunathan.
-Investigated and analyzed the importance of modeling human rationality when learning reward functions from diverse sources of human feedback. This work resulted in a AAAI'23 accepted paper.
-Currently working on training robust models on tasks where features can be spurious in some contexts and reliable in others. Our approach aims to efficiently leverage human input to disambiguate when features are reliable or spurious.

Interpretable Machine Learning for Multimodal Biosensing
University of California, San Francisco *Jun '20 - Aug '21*

-Implemented classification models for multi-modal biosensing datasets.
-Designed and implemented a novel interpretable classification framework using prototype learning for multivariable time series classification.
-Developed simulated datasets for validating the interpretability properties of the framework.
-Wrote a first-author preprint introducing the classification framework.

Deep Reinforcement Learning for Network Scheduling
Lawrence Berkeley National Labs *Jun '19 - Jan '20*

- Conducted research into designing a deep reinforcement learning based agent for scheduling deadline driven data transfers.
- Investigated different state descriptions and reward functions for the reinforcement learning agent.
- Implemented the reinforcement learning agent using TensorFlow and investigated techniques for improving performance.
- Evaluated the reinforcement learning agent in different scenarios, including data transfer requests with different values and heterogeneous link capacities.
- First-authored a paper summarizing the research and results that was accepted in the IFIP Networking 2020 Conference.

Computer Vision Based Plant Phenotyping
University of California, Davis *Jun '18 - Present*

- Worked in Professor Daniel Runcie's Plant Sciences Lab.
- Created a computer vision pipeline for phenotyping a large number of *Arabidopsis thaliana* rosette images using OpenCV. As a result, constructed a dataset containing ∼300,000 growth measurements.
- Designed and applied statistical models for finding Single Nucleotide Polymorphisms (SNPs) associated with variations in the plant growth.
- Used well-established statistical growth models to conduct a genome-wide association study of the plant development using the measurements.
- Investigated environmental influences on *Arabidopsis thaliana* growth using latent variable analysis in R.

Teaching Experience

Teaching Assistant for Introduction to Machine Learning
UC Berkeley EECS Deparment *Jan '22 - May '22*

Served as a 20-hour teaching assistant for CS189/289A, an upper-division/introductory graduate-level machine learning course. Responsible for teaching sections, holding office hours, creating course content, grading exams, and some administrative tasks.

Course Reader for Introduction to Machine Learning
UC Berkeley EECS Deparment *Aug '21 - Dec '22*

Served as a 6-hour course reader for CS189/289A, an upper-division/introductory graduate-level machine learning course. Graded assignments and exams, as well as answering questions during office hours. I also contributed to developing some homework assignments.

Computer Skills

Languages: Python, Java, R, C, RISC-V Assembly
Libraries: PyTorch, Huggingface, TensorFlow, OpenCV, Numpy