Zhaofeng (Philip) Wang

philipwang2025@u.northwestern.edu | Scholar | Github | Website

Education

Northwestern University, BA in Mathematics and Psychology — Sept. 2021 – June 2025

- GPA: 3.9/4.0
- **Relevant Coursework:** Probability and Stochastic Processes, Graduate Seminar in Algorithm, Graduate Seminar in Statistical Network Analysis, Probabilistic Graphical Models, Game Theory and Networked Systems

Research Interest

(Philip: Will come back and make this more coherent) I have a broad interest at the intersection of reinforcement learning, formal methods, and control for safe autonomous systems. I am excited about integrating formal verification and control-theoretic tools with learning-based methods to obtain safe, interpretable, and scalable embodied agents.

Research Experience

IDEAS Lab, Northwestern University, Evanston, IL — Jan. 2024 – Present
Research Assistant, *Advised by Prof. Qi Zhu*

- Built model-based adversarial inverse RL framework with transition-aware reward shaping in stochastic environments, providing theoretical guarantees and substantially improving sample efficiency and performance on stochastic MuJoCo and Atari benchmarks.
- Utilized temporal-logic based verification pipeline to enable the first multi-layer safety evaluation framework for LLM-based embodied agents.
- Developed delay-robust inverse and offline RL frameworks.
- Lead a team of five Master's and undergraduate students to actively develop and maintain a safety benchmark.

Center for Deep Learning, Northwestern University, Evanston, IL — Aug. 2025 – Present
Research Assistant, *Advised by Prof. Zhaoran Wang*

- Proposed a hierarchical latent-token reasoning framework for transformers and empirically validated it on a synthetic dataset with natural language evidence.
- Created a synthetic two-level generative dataset with explicit latent ground truth to perform interpretable circuit-tracing with transcoder-based approach.
- Analyze circuit formation and token-wise behaviors during training; identify reasoning patterns systematically through circuit clustering.

Causal Inference Lab, Northwestern University, Evanston, IL — Mar. 2024 – Aug. 2024
Undergraduate Researcher, *Advised by Prof. Zach Wood-Doughty*

- Augmented benchmark MIMIC-III datasets for medical LLMs in collaboration with physicians to prevent shortcut learning when evaluating clinical reasoning with different NLP tasks within LLMs.
- Fine-tuned open and closed sourced Transformer-based models with LoRA, achieving state-of-the-art results on existing benchmarks and demonstrating a more than 30% performance decrease on the augmented dataset.
- Conducted statistical analysis using scikit-learn to determine the consistency metrics across physicians annotations; quantitatively evaluated the discrepancies between human and LLM annotations.

Publications (* indicates equal contribution) (Philip: Insert paper link later)

SENTINEL: A Multi-Level Formal Framework for Safety Evaluation of LLM-based Embodied Agents. Simon Sinong Zhan*, Yao Liu*, **Philip Wang***, Zinan Wang, Qineng Wang, Zhian Ruan, Xiangyu Shi, Xinyu Cao, Frank Yang, Kangrui Wang, Huajie Shao, Manling Li, Qi Zhu.

Belief-Based Offline Reinforcement Learning for Delay-Robust Policy Optimization. Simon Sinong Zhan, Qingyuan Wu, **Philip Wang**, Frank Yang, Xiangyu Shi, Chao Huang, Qi Zhu

Enhancing Inverse Reinforcement Learning through Encoding Dynamic Information in Reward Shaping. Simon Sinong Zhan*, **Philip Wang***, Qingyuan Wu, Yixuan Wang, Ruochen Jiao, Chao Huang, Qi Zhu.

Inverse Delayed Reinforcement Learning. Simon Sinong Zhan*, Qingyuan Wu*, Aria Ruan, Frank Yang, **Philip Wang**, Yixuan Wang, Ruochen Jiao, Chao Huang, Qi Zhu.

Towards Rigorously Evaluating Clinical Reasoning in LLMs. Christopher Wong, Divy Kumar*, Amiin Muse*, **Philip Wang***, Finn Wintz*, Zach Wood-Doughty

Teaching Experience

Peer Mentor, CS_474 Probabilistic Graphical Models, *Northwestern University* Fall 2024
Academic Advisor, Intensive Law & Trial, *Stanford Law School* Summer 2022, 2023

Leadership Experience

Treasurer and Competing Attorney & Witness, *Northwestern Mock Trial Team* Sept. 2021 - June 2025

- Managed a budget of over $50,000 while serving as treasurer; revised budget structure, leading to a 20% reduction in student membership fee.
- Placed 8th out of ∼700 teams at the National Championship Tournament hosted by the American Mock Trial Association.
- Instructed 120 high school students as advisor throughout six ten-day sessions in the summer about mock trial and law at Stanford Law School, including techniques on research, public speaking, and drafting arguments.

Resident Assistant, *Northwestern Residential Services* Aug. 2022 - June 2025

- Served as an advisor to 150 residents over three years to promote community, individual growth, diversity, and well-being.
- Acted as a liaison between campus organizations and residents, connecting residents to resources as needed in crisis moments and everyday endeavors.
- Managed community funds and led biweekly community events to foster a positive living environment.

Awards

Northwestern Computer Science Department Summer Research Grant	2024
Awarded sponsorship to compete in United States Bridge Championships	2024, 2025
Award for Excellence in Mathematics by a First-Year Student	2022
Selected for US U21 National Bridge Team twice	2019, 2022

Technologies

Languages: Python, C++, Java, Lean

Packages and Frameworks: PyTorch, Scikit-learn, OpenCV, OpenAI Gymnasium, Stable Baselines, AI2Thor, BEHAVIOR-1k