Zihao Wang

920 E 58th St, Office 408, Chicago, IL 60637 • wangzh@uchicago.edu • (312) 394-0229

INTERESTS

- Large Language Model, Text-to-Vision Models, Large-Scale Foundation Modeling

SKILLS

- Expertise in LLM alignment methods using JAX and PyTorch (such as RLHF, DPO, IPO)
- Experience in LLM finetuning, in-context learning, and interpretability methods
- Experience with training large models with massive datasets using data/model parallelism

EDUCATION

University of Chicago | PhD program in Statistics — Sep 2020 - present
University of Chicago | B.S., Computational and Applied Mathematics — Sep 2019

SELECTED PUBLICATIONS

- **Zihao Wang,** Chirag Nagpal, Jonathan Berant, Jacob Eisenstein, Alex D'Amour, Sanmi Koyejo, Victor Veitch (2024). Transforming and Combining Rewards for Aligning Large Language Models (arXiv:2402.00742).
- **Zihao Wang,** Lin Gui, Jeffrey Negrea, Victor Veitch (2023). Concept Algebra for (Score-Based) Text-Controlled Generative Models https://openreview.net/pdf?id=SGlrCuwdsB (**NeurIPS 2023**).
- **Zihao Wang**, Victor Veitch (2022). The Causal Structure of Domain Invariant Supervised Representation Learning (arXiv:2208.06987).
- Peter Carbonetto, Abhishek Sarkar, **Zihao Wang**, Matthew Stephens (2021), Non-negative matrix factorization algorithms greatly improve topic model fits (arXiv:2105.13440).

WORK EXPERIENCE

Google Deepmind (Student Researcher**)** — June 2023-Jan 2024
Manager: **Prof. Sanmi Koyejo**, Stanford Computer Science, Google Deepmind

- Siginifcantly improved the Reinforcement Learning from Human Feedback (RLHF) pipeline for Large Language Model alignment by reward transformation and aggregation. Paper submitted to ICML.
- Motivated by the goal to interpret reward models as probabilistic quantities, which led to the innovation of using prompt-specific baselines for clearer model interpretation.
- Following this motivation, implemented strategic modifications to the RLHF pipeline to effectively mitigate reward over-optimization and improve reward aggregation.
- Achieving better win-rate this with less than one-third of the original KL budget.

RESEARCH PROJECTS

Concept Algebra for (Score-Based) Text-Controlled Generative Models — Oct 2022 - May 2023
Advisor: **Prof. Victor Veitch**, Department of Statistics University of Chicago, Google Deepmind

- Motivated by the failure modes of text-to-image models for correlated concepts, we delved into the study of internal representations of latent concepts, aiming to identify and manipulate these concepts in isolation during the generation process.
- Theorized and formalized how concepts are encoded as subspaces within a designated representation space, leading to a new understanding of concept-specific encodings.
- Developed an innovative method for altering the model's expressed concepts through algebraic manipulation of these representations, enhancing model interpretability and control.
- Demonstrated the efficacy of concept algebra in experiments by disentangling rare concept combinations, such as creating images of unusual subject/style pairings, showcasing the method's superiority over direct prompting and traditional concept composition techniques.

The Causal Structure of Domain Invariant Supervised Representation Learning — Jan 2022-Sep 2022
Advisor: **Prof. Victor Veitch**, Department of Statistics University of Chicago, Google Deepmind

- Studied the problem of Domain Shifts in Machine Learning through a novel Causal Framework
- Characterized the relationships among various domain-invariant representation learning methods: data augmentation, distributional-invariance learning and invariant risk minimization, and give recommendations for which methods to use in practice
- Performed experiments on synthetic and large-scale problems with domain shifts with Pytorch

Nonnegative Matrix Factorization (NMF) on count data — June 2018 – August 2020
Advisor: **Prof. Matthew Stephens**, Department of Statistics and Human Genetics, University of Chicago

- Proposed Empirical Bayes approach to nonnegative matrix factorization for count data
- Implemented methods in in R package ebpmf, and apply to large-scale genetics and text datasets, showing improvement in interpretability