Zihao **Wang**

PhD Candidate in Statistics at the University of Chicago

☎ (+1) 312-394-0229 | ✉ wangzh@uchicago.edu | in wzihao12 | 🎓 Zihao Wang

Summary

AI researcher with statistics background, specializing in safety and alignment for large generative models (LLMs, Diffusion Models). Published in top-tier conferences including ICML and NeurIPS. Interned at Google DeepMind and ByteDance LLM Security team. PhD candidate in UChicago Statistics, developing principled methods for responsible AI inspired by Bayesian and causal inference approaches.

Education

University of Chicago
Chicago, IL

PhD candidate in Statistics (Advisor: Victor Veitch) *Sep 2020 - present*

University of Chicago
Chicago, IL

B.S., Computational and Applied Mathematics *Sep 2019*

Selected Publications

Transforming and Combining Rewards for Aligning Large Language Models (**ICML 2024**) *2024*

- Authors: **Zihao Wang**, Chirag Nagpal, Jonathan Berant, Jacob Eisenstein, Alex D'Amour, Sanmi Koyejo, Victor Veitch
- TLDR: A reward transformation technique for LLM RLHF significantly alleviates reward hacking and improves reward aggregation.

Concept Algebra for (Score-Based) Text-Controlled Generative Models (**NeurIPS 2023**) *2023*

- Authors: **Zihao Wang**, Lin Gui, Jeffrey Negrea, Victor Veitch
- TLDR: A method for identifying concept subspaces in text-guided diffusion models enables algebraic manipulation of representations, allowing for precise control over entangled concepts with theoretical guarantees.

PFT: Enhancing Prompt Injection Robustness via Position-Enhanced Finetuning *2024*

- Authors: **Zihao Wang**, Yibo Jiang, Jiaohao Yu, Heqing Huang
- TLDR: A new position-enhanced fine-tuning approach improves LLMs' ability to distinguish between system instructions and user input, enhancing robustness against prompt injection attacks without compromising performance.

Does Editing Provide Evidence for Localization? (ICML 2024, Mechanistic Interpretability workshop) *2024*

- Authors: **Zihao Wang**, Victor Veitch
- TLDR: A critical examination of LLM interpretability methods reveals that localized edits causing targeted behavior changes may not indicate true behavior localization, challenging current interpretability methodologies in the field.

The Causal Structure of Domain Invariant Supervised Representation Learning (ICML 2022, SCIS workshop) *2022*

- Authors: **Zihao Wang**, Victor Veitch
- TLDR: An analysis of "invariant" representation learning methods for domain shift reveals their effectiveness depends critically on the underlying causal structure of data, clarifying when these approaches can yield robust models.

Non-negative matrix factorization algorithms greatly improve topic model fits *2021*

- Authors: Peter Carbonetto, Abhishek Sarkar, **Zihao Wang**, Matthew Stephens

Work Experience

ByteDance (LLM Security Team)
San Jose, CA

Research Intern (Manager: Xinyu Xing & Heqing Huang, ByteDance) *July 2024 - Sep 2024*

- Finetuned close-domain LLMs to address critical security vulnerabilities, achieving >90% reduction in attack success rate.
- Developed system prompt leakage detectors for LLM bots, achieving >50% improvement in detection rate over baselines in real traffic.
- Contributed to an AI threat database initiative, performing feature engineering and hierarchical clustering.

Google DeepMind
Remote

Student Researcher (Manager: Sanmi Koyejo, Stanford University, Google DeepMind) *June 2023 - Jan 2024*

- Developed novel reward transformation and aggregation techniques, resulting in significant improvements over baseline RLHF.
- Innovated prompt-specific reward baselines for clearer model interpretation.
- Achieved substantial performance gains: significantly improved win-rates and reduced reward hacking compared to baseline RLHF.

Skills

AI Expertise

- Deep Learning, NLP, LLM, RLHF, SFT, Diffusion Models, Prompt Engineering, LLM Interpretability, Causal Inference, Bayesian Modeling

Programming

- Python, R, PyTorch, JAX, TensorFlow, Hugging Face Libraries, Scikit-learn, Pandas, NumPy, Model/Data sharding, GCP