Yihao Zhang

China, Beijing jekyllzhang@gmail.com zhangyihao.site Zhang-Yihao

Education

B.S. in Data Science and Big Data Technology Sept 2020 – Jul 2024
School of Mathematical Sciences, Peking University
Ph.D. in Applied Mathematics Sept 2024 –
School of Mathematical Sciences, Peking University

Research Interests

- Mechanistic Interpretability for Large Language Models
- Causality in AI, Formalization and Verification of Causality Related Issues
- Large Language Model Alignment, Trustworthy LLM
- Representation Engineering in LLM
- AI Safety, Verifying Robustness/Fairness/Trustworthy Related Issues in AI
- Automated Interactive Theorem Proving, AI4ITP
- Formal Methods, Model Checking, Software Analysis, Program Verification
- Formalizing and Verifying Quantum Computation Systems and Quantum AI Systems
- Testing Technologies for AI System

Selected Publications

On the Duality Between Sharpness-Aware Minimization and Adversarial Training **ICML 2024**
Yihao Zhang*, Hangzhou He*, Jingyu Zhu*,
Huanran Chen, Yifei Wang, Zeming Wei†

Towards General Conceptual Model Editing via Adversarial Representation Engineering **NeurIPS 2024**
Yihao Zhang, Zeming Wei, Jun Sun†, Meng Sun†

Weighted Automata Extraction and Explanation of Recurrent Neural Networks for Natural Language Tasks **JLAMP, Vol 136**
Zeming Wei, Xiyue Zhang, **Yihao Zhang**, Meng Sun†

MILE: A Mutation Testing Framework of In-Context Learning Systems **SETTA 2024**
Zeming Wei, **Yihao Zhang**, Meng Sun†

MedTiny: Enhanced Mediator Modeling Language for Scalable Parallel Algorithms **QRS-C 2023**
Xiangyu Li, **Yihao Zhang**, Xiaokun Luan, Xiaoyong Xue, Meng Sun†

Using Z3 for Formal Modeling and Verification of FNN Global Robustness **SEKE 2023**
Yihao Zhang, Zeming Wei, Xiyue Zhang, Meng Sun†

Research Experience

S^4 **Group Member, Peking University** Apr 2022 –
Advisor: Meng Sun

- Formalizing and Verifying Smart Contracts
- Automata Extraction from AI Models
- Verification of Neural Network Robustness

Research Assistant, Singapore Management University Oct 2023 – May 2024
Advisor: Jun Sun

- Representation Engineering in LLMs
- Aligning Large Language Models via Model Editing

Projects

Study on the interpretability of large language model architecture and algorithm 2023-2025
Program Director

Trustworthy guarantee of deep learning system 2022-2025
Program Major Member

Honors and Awards

- Huaixin Bachelor (Honours Degree), 2024
- Selected for the Elite Program (Graduate) in the School of Mathematical Sciences, Peking University.
- University Scholarship (School of Mathematical Sciences), Peking University, 2023
- Second prize, Chinese Mathematics Competitions for Undergraduates (Beijing Division), 2023

Other Manuscripts

Sharpness-aware Minimization Alone Can Improve Adversarial Robustness **ICML 2023 AdvML-Frontiers Workshop**
Zeming Wei*†, Jingyu Zhu*, **Yihao Zhang***

Boosting Jailbreak Attack with Momentum **ICLR 2024 R2-FM Workshop**
Yihao Zhang*, Zeming Wei*†

Exploring the Robustness of In-Context Learning with Noisy Labels **ICLR 2024 R2-FM Workshop**
Chen Cheng*, Xinzhi Yu*, Haodong Wen*, Jinsong Sun, Guanzhang Yue, **Yihao Zhang**, Zeming Wei†

Automata Extraction from Transformers **Arxiv Preprint**
Yihao Zhang, Zeming Wei, Meng Sun†