Wei Yao

Ph.D. Student, Renmin University of China

Expected graduation: 2027 | Beijing, China

wei.yao@ruc.edu.cn | Homepage: yw101004244.github.io | Google Scholar

Research Interests

I study how large language models learn under imperfect supervision. My recent work focuses on generalization in teacher-student learning, particularly weak-to-strong generalization and on-policy distillation. I combine theory and experiments to understand how pre-training and post-training shape this process and develop methods to improve model performance. My earlier work studied LLM interpretability, as well as robustness and fairness in deep learning.

Education

Renmin University of China Sep 2022–Present
Ph.D. Student, Gaoling School of Artificial Intelligence
Advisor: Prof. Yong Liu

Huazhong University of Science and Technology Sep 2018–Jun 2022
B.Eng. in Software Engineering
Advisor: Prof. Kun He

Research Experience

National University of Singapore Nov 2025–Aug 2026
Visiting Ph.D. Student
Host: Prof. Yunbei Xu

Shanghai Artificial Intelligence Laboratory Oct 2023–Mar 2024
Research Intern
Mentor: Dr. Jing Shao

Teaching Experience

Introduction to Deep Learning Fall 2022
Teaching Assistant (Undergraduate Course), Renmin University of China
Instructor: Prof. Yong Liu
Prepared exam questions, graded assignments, and answered students' questions.

Honors

Outstanding Reviewer 2025
EMNLP 2025

Merit Student 2023–2026
Renmin University of China

Merit Student 2019–2022
Huazhong University of Science and Technology

National Scholarship 2019
Huazhong University of Science and Technology

Academic Service

Reviewer: ICML, NeurIPS, ICLR, ACL, EMNLP, CVPR, ICCV, ECCV, AAAI, TMLR, AISTATS.

Selected Publications

* Equal contribution.

[1] **Weak-to-Strong Generalization via Bregman Bias-Variance Decomposition**
Gengze Xu*, **Wei Yao***, Ziqiao Wang, Yong Liu
ICML 2026

[2] **Revisiting Weak-to-Strong Generalization in Theory and Practice: Reverse KL vs. Forward KL**
Wei Yao*, Wenkai Yang*, Ziqiao Wang, Yankai Lin, Yong Liu
Findings of ACL 2025

[3] **Understanding Model Ensemble in Transferable Adversarial Attack**
Wei Yao*, Zeliang Zhang*, Huayi Tang, Yong Liu
ICML 2025

[4] **Understanding Fairness Surrogate Functions in Algorithmic Fairness**
Wei Yao*, Zhanke Zhou*, Zhicong Li, Bo Han, Yong Liu
TMLR 2024 Presented at ICLR 2025 (Journal-to-Conference Track)

[5] **Towards Tracing Trustworthiness Dynamics: Revisiting Pre-training Period of Large Language Models**
Chen Qian*, Jie Zhang*, **Wei Yao***, Dongrui Liu, Zhenfei Yin, Yu Qiao, Yong Liu, Jing Shao
Findings of ACL 2024

[6] **Random Smooth-based Certified Defense against Text Adversarial Attack**
Zeliang Zhang*, **Wei Yao***, Susan Liang, Chenliang Xu
Findings of EACL 2024

[7] **Fair Scratch Tickets: Finding Fair Sparse Networks without Weight Training**
Pengwei Tang*, **Wei Yao***, Zhicong Li, Yong Liu
CVPR 2023

Preprints

[1] **On the Blessing of Pre-training in Weak-to-Strong Generalization**
Wei Yao, Wang Zhaoyang, Gengze Xu, Chen Qian, Dongrui Liu, Ziqiao Wang, Yong Liu, Yunbei Xu
arXiv:2605.05710 2026

[2] **On Weak-to-Strong Generalization and f-Divergence**
Wei Yao*, Gengze Xu*, Huayi Tang, Wenkai Yang, Donglin Di, Ziqiao Wang, Yong Liu
arXiv:2506.03109 2025