Jihan Yao

Ph.D student at Paul G. Allen School of CSE, University of Washington
(206)-492-3993 | jihany2@cs.washington.edu | https://yaojh18.github.io/

Research Interest

My research interests primarily focus on Large Foundation Models:

- **Reliability**: Investigate whether models can recognize their knowledge gaps and abstain when uncertain. This enhances trustworthiness and minimizes risks in high-stakes domains such as medical applications.
- **Evaluation**: Address biases in model evaluation and align with human preferences, particularly in multi-modal tasks. This will establish standardized, fair, and reliable evaluation protocols for large foundation models.
- **Alignment**: Explore data-centric approaches to optimize the alignment process. This can expand the capabilities of large foundation models under only limited high-quality data and potentially enable self-improvement.

Education

University of Washington — Seattle, WA, USA
Ph.D student at Paul G. Allen School of Computer Science & Engineering — Sep. 2023 – present
Advisor: Banghua Zhu

Tsinghua University — Beijing, CN
B.S. at Department of Computer Science and Technology — Sep. 2018 – Jun. 2023
GPA: 3.95 / 4.00, Ranking: 7 / 210

Selected Research Experience

Characterizing Abstention Behavior in Vision Language Models — University of Washington
Research Assistant, advised by Lucy Lu Wang and Banghua Zhu — Oct. 2024 – Present

- Increasing attention in VLMs has been shifted to visual reasoning tasks such as MMMU, which require capabilities of both image recognition and reasoning with parameterized knowledge.
- Developed a knowledge-gap-aware dataset and benchmarked existing VLMs with abstention mechanisms.
- While self-consistency is the best, VLM-as-a-judge fails due to misalignment between visual and textual modalities.

Aligning Large Language Models with Wrong Answers Only — University of Washington
Research Assistant, advised by Lucy Lu Wang and Yulia Tsvetkov — May 2024 – Oct. 2024

- LLMs may face challenges in tasks where reliable or cost-effective ground-truths are unavailable.
- Proposed that wrongness is a spectrum and aligned LLMs to prefer less wrong answers over more wrong ones.
- Experiments show that LLM-as-a-judge provides reliable wrong-over-wrong preferences. Aligning with only wrong answers improves model calibration, reducing wrongness by up to 9.0%, and increasing correct answers by 7.0%.

Publications / Preprint

- **Jihan Yao***, Wenxuan Ding*, Shangbin Feng*, Lucy Lu Wang, Yulia Tsvetkov. *Varying Shades of Wrong: Aligning LLMs with Wrong Answers Only.*
In submission to ICLR 2025

- Bingbing Wen, **Jihan Yao**, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, Lucy Lu Wang. *Know Your Limits: A Survey of Abstention in Large Language Models.*
Conditional Acceptance to TACL 2024

- Yuta Saito, **Jihan Yao**, Thorsten Joachims. *POTEC: Off-Policy Learning for Large Action Spaces via Two-Stage Policy Decomposition.*
In submission to ICLR 2025

- Maryam Amirizaniani, **Jihan Yao**, Adrian Lavergne, Elizabeth Snell Okada, Aman Chadha, Tanya Roosta, Chirag Shah. *LLMAuditor: A Framework for Auditing Large Language Models Using Human-in-the-Loop.*
Preprint

Awards

Excellent Comprehensive Scholarship of Tsinghua University 2020

Excellent Academic Scholarship of Tsinghua University 2019, 2021

Services

- **Reviewer**: NeurIPS (2024), ICLR (2025), ACL (2025)
- **Teaching Assistant**: CSE344: Introduction to Data Management (Spring 2024), CSE414: Introduction to Database Systems (Fall 2024)

Skills

- **Programming Languages**: Python, C/C++, Java, JavaScript
- **Machine Learning**: Pytorch, Huggingface
- **Language**: English (TOEFL iBT 112), Chinese (native)