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:

- **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.
- **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 unified, fair, and reliable evaluation protocols for large foundation models.

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

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 Score: 8666
- 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 Score: 8886
- 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

Selected Research Experience

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 are challenged by tasks where reliable or cost-effective ground-truths are unavailable.
- How to effectively use low-quality data remains underexplored. 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 can make models generate more correct answers by 7.0% and improve calibration.

Improve Abstention Capability for Vision Language Models University of Washington
Research Assistant, advised by Lucy Lu Wang and Banghua Zhu Oct. 2024 – Present

- State-of-the-art abstention mechanism like multi-agents cooperation or consistency check failed for VLMs.
- Found that VLMs are significantly over-confident about their image perception.

- Proposed a two-stage calibration on both image perception and answer accuracy.

Verifiable Instruction Following Evaluation for Any-to-Any Models University of Washington
Research Assistant, advised by Banghua Zhu Oct. 2024 – Present

- Any-to-any model evaluation benchmarks usually employ MLLM-as-a-judge, proving to be misaligned with human preferences by recent studies.
- A verifiable instruction dataset is curated by carefully designing prompts for MLLMs or employing self-evident evaluation tools, ensuring at least 80% agreement (48.2% for best previous work) with human annotated labels.
- On top of that, high-quality seed prompts comparable with single modality evaluation benchmarks, unified and compositional multi-modal evaluation are guaranteed.

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)