Binwei Yao

binwei.yao@wisc.edu | LinkedIn | Google Scholar | Homepage | Madison, WI

Research Interests

Post-Training: Alignment, Personalization, Multilinguality
Trustworthy AI: Safety, Fairness, Human-AI Alignment
AI Agent: Multi-agent Collaboration, Reasoning

Education

University of Wisconsin-Madison, Madison, WI — Sept. 2023 – Present
Ph.D. in Computer Sciences

Shanghai Jiao Tong University, Shanghai, China — Sept. 2018 – Jun. 2022
B.Eng. in Software Engineering

Industry Experience

Amazon AWS Agentic AI — Seattle, WA — May. 2025 – Aug. 2025
Applied Scientist Intern, Working on Agent Safety

Publications

No Preference Left Behind: Group Distributional Preference Optimization
Binwei Yao, Zefan Cai, Yun-Shiuan Chuang, Shanglin Yang, Ming Jiang, Diyi Yang, Junjie Hu
*Accepted by **ICLR 2025*** [Paper]

Benchmarking Machine Translation with Cultural Awareness
Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, Junjie Hu
*Accepted by **EMNLP 2024** (Findings)* [Paper]

D^4: a Chinese Dialogue Dataset for Depression-Diagnosis-Oriented Chat
Binwei Yao, Chao Shi, Likai Zou, Lingfeng Dai, Lu Chen, Mengyue Wu, Zhen Wang, Kai Yu.
*Accepted by **EMNLP 2022** (Oral)* [Paper]

MSDWild: Multi-modal Speaker Diarization Dataset in the Wild
Tao Liu*, Shuai Fan*, Xu Xiang, Hongbo Song, Shaoxiong Lin, Jiaqi Sun, Tianyuan Han, Siyuan Chen, **Binwei Yao**, Sen Liu, Yifei Wu, Yanmin Qian and Kai Yu.
*Accepted by **Interspeech 2022*** [Paper]

Preprints

Peacemaker or Troublemaker: How Sycophancy Shapes Multi-Agent Debate
Binwei Yao, Shang Chao, Wanyu Du, Jianfeng He, Hang Su, Yi Zhang, Sandesh Swamy, Yanjun Qi
Submitted to ICLR 2026 [Paper]

DEBATE: A Large-Scale Benchmark for Role-Playing LLM Agents in Multi-Agent, Long-Form Debates
Yun-Shiuan Chuang, Ruixuan Tu, Chengtao Dai, Smit Vasani, **Binwei Yao**, Michael Henry Tessler, Sijia Yang, Dhavan V. Shah, Robert D. Hawkins, Junjie Hu, Timothy T. Rogers
Submitted to ICLR 2026 [Paper]

AI as a Deliberative Partner Fosters Intercultural Empathy for Americans but Fails for Latin American Participants
Isabel Villanueva, Tara Bobinac, **Binwei Yao**, Junjie Hu, Kaiping Chen
Under Review [Paper]

Towards Reliable and Empathetic Depression-Diagnosis-Oriented Chats

Kunyao Lan, Cong Ming, **Binwei Yao**, Lu Chen, Mengyue Wu
Under Review [Paper]

Research Experiences

HuLab, **University of Wisconsin-Madison** Madison, WI Sept. 2023 – Present
Group Distributional Alignment for LLMs
Advisor: Prof. Junjie Hu

SALT Lab, **Stanford University** Remote Oct. 2022 – June. 2023
Culture-Aware Machine Translation System
Advisors: Prof. Diyi Yang, Prof. Junjie Hu

X-LANCE Lab, Shanghai Jiao Tong University Shanghai, China Dec. 2020 – June. 2023
The Dialogue System for Depression Diagnosis
Advisors: Prof. Mengyue Wu, Prof. Lu Chen, Prof. Kai Yu

Skills

Programming Languages and Tools: *Python*, *C++*, *Java*, *Cuda*, *Go*; *PyTorch* and *LaTeX*