Peixuan Han

University of Illinois Urbana-Champaign • Illinois, U.S.
ph16@illinois.edu

Education

University of Illinois Urbana-Champaign — Illinois, U.S.
Ph.D., Siebel School of Computing and Data Science. — *2024.8 - Present*

- **Advisor:** Prof. Jiaxuan You.
- **Research Interests:** Large language model safety & reasoning, especially social intelligence.

Carnegie Mellon University — Pennsylvania, U.S.
Visiting Student, School of Computer Science — *2023.6 - 2023.9*

Tsinghua University — Beijing, China
Undergraduate, Department of Computer Science and Technology. — *2020.9 - 2024.6*

- Attended college at 15.
- **GPA**: 3.91/4.00
- AI-relevant courses: Introduction to AI, Data Structure, Artificial Networks, Information Retrieval.

Publications

Peixuan Han, Zijia Liu, Jiaxuan You. ToMAP: Training Opponent-Aware LLM Persuaders with Theory of Mind. *preprint version*.

Peixuan Han, Cheng Qian, Xiusi Chen, Yuji Zhang, Denghui Zhang, Heng Ji. Internal Activation as the Polar Star for Steering Unsafe LLM Behavior. *preprint version*.

Peixuan Han, Zhenghao Liu, Zhiyuan Liu, Chenyan Xiong. Enhancing Dense Retrievers' Robustness with Group-level Reweighting. *preprint version*.

Alexi Gladstone, Ganesh Nanduru, Md Mofijul Islam, **Peixuan Han**, Hyeonjeong Ha, Aman Chadha, Yilun Du, Heng Ji, Jundong Li, Tariq Iqbal. Energy-Based Transformers are Scalable Learners and Thinkers. *preprint version*.

Zijia Liu, **Peixuan Han**, Haofei Yu, Haoru Li, Jiaxuan You. Time-R1: Towards Comprehensive Temporal Reasoning in LLMs. *preprint version*.

Xiusi Chen, Shanyong Wang, Cheng Qian, Hongru Wang, **Peixuan Han**, Heng Ji. DecisionFlow: Advancing Large Language Model as Principled Decision Maker. *preprint version*.

Kunlun Zhu, Jiaxun Zhang, Ziheng Qi, Nuoxing Shang, Zijia Liu, **Peixuan Han**, Yue Su, Haofei Yu, Jiaxuan You. SafeScientist: Toward Risk-Aware Scientific Discoveries by LLM Agents. *preprint version*.

Cheng Qian, **Peixuan Han**, Qinyu Luo, Bingxiang He, Xiusi Chen, Yuji Zhang, Hongyi Du, Jiarui Yao, Xiaocheng Yang, Denghui Zhang, Yunzhu Li, Heng Ji. EscapeBench: Pushing Language Models to Think Outside the Box (ACL'25). *preprint version*.

Yanci Liu, Jiayu Li, **Peixuan Han**, Siyu Ma, Feng Du, Min Zhang. Emotion and Passage of Time Judgment: Evidence from Weibo Dataset, *under review*.

Awards

- **Golden Prize**, International Junior Science Olympiad 2018.11
- **First Prize**, Asia-Pacific Informatics Olympiad 2019.5
- **Scholarship for Academic Excellence**, Tsinghua University, 2021.9, 2022.9
- **Scholarship for Social Work**, Tsinghua University, 2022.9
- **Outstanding Graduate of CS Department**, Tsinghua University. 2024.6

Skills

English Ability: TOEFL - Total 119, Reading 30, Writing 30, Listening 30, Speaking 29; GRE - Verbal 164, Quantitative 170, Writing 4.0.

Programming: Proficient in Python, C++, SQL. Proficient in data structures and algorithms.

Deep Learning Framework: Proficient in PyTorch, Huggingface Transformers, verl.

Service

- **Teaching Assistant**, Introduction to Artificial Intellligence 2023.9-2024.1
- **Reviewer**, COLING'25 2024.11
- **Reviewer**, KDD'25, Structured Knowledge in LLMs Workshop 2025.4