Jiaxin Wen

wenjx22@mails.tsinghua.edu.cn | jiaxin-wen.github.io | github/Jiaxin-Wen

Education

Tsinghua University Beijing, China
MA in Computer Science and Technology *2022 - Present*
- GPA: 3.92/4.00
- Advisor: Prof. Minlie Huang

Tsinghua University Beijing, China
B.Eng. in Computer Science and Technology *2018 - 2022*
- GPA: 3.80/4.00
- Advisor: Prof. Minlie Huang

Honors and Awards

Research Excellence Award, Tsinghua University	*2023*
Global AI Innovation Contest (6nd out of 5,000)	*2022*
Outstanding Graduate, Dept. CST, Tsinghua University	*2022*
Outstanding Undergraduate Thesis, Tsinghua University	*2022*
Global AI Innovation Contest Runner-up (2nd out of 5,000)	*2021*
Academic Excellence Award, Tsinghua University	*2020*
Volunteer Excellence Award, Tsinghua University	*2020*
Volunteer Excellence Award, Tsinghua University	*2019*
Philobiblion Award, Tsinghua University	*2019*

Research

Balacing Safety and Usefulness in the Long-term Deployment of Untrusted LLMs
ICLR2025 submission
Jiaxin Wen*, Caleb Larson*, Vivek Hebbar*, Aryan Bhatt, Ansh Radhakrishnan, Mrinank Sharma, Henry Sleight, He He, Shi Feng, Ethan Perez, Buck Shlegreis, Akbir Khan

Language Models Learn to Mislead Humans via RLHF
ICLR2025 submission
Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R. Boman, He He, Shi Feng

Unlocking Reasoning Potential in Large Language Models by Scaling Code-form Planning
ICLR2025 submission
Jiaxin Wen*, Jian Guan*, Hongning Wang, Wei Wu, Minlie Huang

Backdoored LLM Agents that Detect Human Overseers
ICML2024 Workshop
Heng Wang, Ruiqi Zhong, **Jiaxin Wen**, Jacob Steinhardt

Learning Task Decomposition to Assist Humans in Competitive Programming
ACL2024
Jiaxin Wen, Ruiqi Zhong, Pei Ke, Zhihong Shao, Minlie Huang

Unveiling the Implicit Toxicity in Large Language Models
EMNLP2023
Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, Minlie Huang

ETHICIST: Targeted Training Data Extraction Through Loss Smoothed Soft Prompting and Calibrated Confidence Estimation
ACL2023(oral)
Zhexin Zhang, **Jiaxin Wen**, Minlie Huang

Re^{3}Dial: Retrieve, Reorganize and Rescale Dialogue Corpus for Long-Turn Open-Domain Dialogue Pre-training
EMNLP2023
Jiaxin Wen, Hao Zhou, Jian Guan, Minlie Huang

AugESC: Large-Scale Data Augmentation for Emotional Support Conversation with Pre-trained Language Models
ACL2023 findings
Chujie Zheng, Sahand Sabour, **Jiaxin Wen**, Minlie Huang

AutoCAD: Automatically Generate Counterfactuals for Mitigating Shortcut Learning
EMNLP2022 findings
Jiaxin Wen, Yeshuang Zhu, Jinchao Zhang, Jie Zhou, Minlie Huang

EVA2.0: Investigating Open-Domain Chinese Dialogue Systems with Large-Scale Pre-Training
Machine Intelligence Research
Yuxian Gu*, **Jiaxin Wen***, Hao Sun*, Yi Song, Pei Ke, Chujie Zheng, Zheng Zhang, Jianzhu Yao, Xiaoyan Zhu, Jie Tang, Minlie Huang

Persona-Guided Planning for Controlling the Protagonist's Persona in Story Generation
NAACL2022(oral)
Zhexin Zhang*, **Jiaxin Wen***, Jian Guan, Minlie Huang

Robustness Testing of Language Understanding in Task-Oriented Dialog
ACL2021(Oral)
Jiexi Liu*, Ryuichi Takanobu*, **Jiaxin Wen**, Dazhen Wan, Hongguang Li, Weiran Nie, Cheng Li, Wei Peng, Minlie Huang

Projects

ChatGLM3 Code Interpreter
- We implement a tool-augmented large language model that can seamlessly interact with python interpreter to solve challenging tasks.
- https://chatglm.cn/main/code

Ai-Topia
- We implement a Chinese open-domain character conversational agent
- Since its launch on 2022/11/15, Ai-Topia has received more than **1,000,000** calls and **40,000** unique users.
- Chat with Ai-Topia on WeChat!

OPD
- The largest open-source Chinese open-domain pre-trained dialogue model (6.3B parameters)
- Blog: http://coai.cs.tsinghua.edu.cn/static/opd/posts/opd_blog/
- Github: https://github.com/thu-coai/OPD

Emohaa
- A Chinese empathetic chatbot based on large-scale pre-trained models.
- Since its launch in March 2022, Emohaa has received more than **520, 000** calls.
- Chat with Emohaa on WeChat!

EVA
- A large-scale open-source Chinese dialogue pre-trained model (2.8B parameters)
- Our code and pre-trained models are publicly available at https://github.com/thu-coai/EVA

Service

Reviewer

- 2024: ACL (Safety, LLM for Programming, Dialogue), COLM (Safety)
- 2023: EMNLP (Dialogue, Safety)
- 2023: ACL (Large-scale Pre-training)
- 2022: EMNLP (Dialogue and Interactive Systems)

Experience

Anthropic San Francisco, US
Research Contractor *2024.6-*

Alignment Research Group, NYU New York, US
Research Intern *2024.4-*

Language Model For Reasoning Team, Ant Research Beijing, China
Research Intern *2024.3-2024.9*

Foundational Model Team, Zhipu AI Beijing, China
Research Intern *2023.7 - 2023.11*

WeChat AI Team, Tencent Beijing, China
Research Intern *2021.6 - 2021.12*

WeChat AI Team, Tencent Beijing, China
Algorithm Intern *2020.6 - 2020.10*