Chujie Zheng 郑楚杰

Contact Information

Address	Room 4-504, FIT Building, Tsinghua University, Beijing 100084, China
Email	chujiezhengchn@gmail.com
Homepage	https://chujiezheng.github.io/
Google Scholar	https://scholar.google.com/citations?user=55zBNgUAAAAJ

Education

- **Ph.D. student**, Department of Computer Science and Technology, Tsinghua University — Sep 2020 – Present
  Advisor: Minlie Huang
- **Visiting Scholar**, Computer Science Department, UCLA — Nov 2023 – Present
  Advisor: Nanyun (Violet) Peng
- **B.S.** in Physics, Tsinghua University — Aug 2016 – Jun 2020

Research Overview and Highlights

- I have a broad research interest in **building trustworthy AI systems**, with the current focus on **LLM alignment and safety**. My recent researches made efforts to understand the intrinsic working mechanisms of modern LLMs (preprint, ICLR 2024 Spotlight), discover their limitations and risks (preprint, ICLR 2024 Spotlight), and improve them to align with human expectations and social norms (preprint, ACL 2023 Findings). Specifically:
  - The ACL 2023 paper shows we can train LLMs to directly align sequence likelihood (i.e., generation probability) with reward models (neural classifiers or rule-based functions) to achieve notable controllability.
  - The ICLR 2024 paper investigates LLMs' evaluation bias and debiasing approach in MCQ benchmarks.
  - The preprint paper investigates the working mechanisms of safety prompts in safeguarding LLMs.
  - Devoting all passion and wistom, my ultimate research goal is to *advance and oversee AI systems in a scalable, automated way (with minimal human intervention) and ensure they operate transparently and responsibly*.
- Previously, I also conducted a series of research on **LLMs for social good**, covering the topics of emotional support dialogue systems (ACL 2021, ACL 2023) and empathetic dialogue generation (ACL 2021, AAAI 2022, ACL 2023).
- In my early research career, I built a series of **popular NLP datasets**, including ChID (ACL 2019), KDConv (ACL 2020), ESConv (ACL 2021), CDConv (EMNLP 2022), DiaSafety (ACL 2022), and COLD (EMNLP 2022).
- Google Scholar citations **840**+, h-index **14**, i10-index **16**.

Selected Papers (*: Equal Contribution)

1. **Chujie Zheng**, Fan Yin, Hao Zhou, Fandong Meng, Jie Zhou, Kai-Wei Chang, Minlie Huang, Nanyun Peng. On *Prompt-Driven Safeguarding for Large Language Models*. arXiv:2401.18018.
2. **Chujie Zheng**, Hao Zhou, Fandong Meng, Jie Zhou, Minlie Huang. *Large Language Models Are Not Robust Multiple Choice Selectors*. ICLR 2024 **(Spotlight: 5%)**.
3. **Chujie Zheng**, Pei Ke, Zheng Zhang, Minlie Huang. *Click: Controllable Text Generation with Sequence Likelihood Contrastive Learning*. ACL 2023 Findings.
4. **Chujie Zheng**, Sahand Sabour, Jiaxin Wen, Zheng Zhang, Minlie Huang. *AugESC: Dialogue Augmentation with Large Language Models for Emotional Support Conversation*. ACL 2023 Findings.
5. Siyang Liu*, **Chujie Zheng***, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, Minlie Huang. *Towards Emotional Support Dialog Systems*. ACL 2021.

Other Papers

6. Jinfeng Zhou*, **Chujie Zheng***, Bo Wang, Zheng Zhang, Minlie Huang. *CASE: Aligning Coarse-to-Fine Cognition and Affection for Empathetic Response Generation*. ACL 2023.
7. **Chujie Zheng***, Jinfeng Zhou*, Yinhe Zheng, Libiao Peng, Zhen Guo, Wenquan Wu, Zhengyu Niu, Hua Wu, Minlie Huang. *CDConv: A Benchmark for Contradiction Detection in Chinese Conversations*. EMNLP 2022.
8. Jiawen Deng*, Jingyan Zhou*, Hao Sun, **Chujie Zheng**, Fei Mi, Helen Meng, Minlie Huang. *COLD: A Benchmark for Chinese Offensive Language Detection*. EMNLP 2022.
9. Hao Sun*, Guangxuan Xu*, Jiawen Deng, Jiale Cheng, **Chujie Zheng**, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, Minlie Huang. *On the Safety of Conversational Models: Taxonomy, Dataset, and Benchmark*. ACL 2022 Findings.
10. Sahand Sabour, **Chujie Zheng**, Minlie Huang. *CEM: Commonsense-aware Empathetic Response Generation*. AAAI 2022.
11. **Chujie Zheng**, Minlie Huang. *Exploring Prompt-based Few-shot Learning for Grounded Dialog Generation*. arXiv:2109.06513.
12. Hao Zhou*, Pei Ke*, Zheng Zhang*, Yuxian Gu, Yinhe Zheng, **Chujie Zheng**, Yida Wang, Chen Henry Wu, Hao Sun, Xiaocong Yang, Bosi Wen, Xiaoyan Zhu, Minlie Huang, Jie Tang. *EVA: An Open-Domain Chinese Dialogue System with Large-Scale Generative Pre-Training*. arXiv:2108.01547.
13. **Chujie Zheng**, Yong Liu, Wei Chen, Yongcai Leng, Minlie Huang. *CoMAE: A Multi-factor Hierarchical Framework for Empathetic Response Generation*. ACL 2021 Findings.
14. Hao Sun*, Zhenru Lin*, **Chujie Zheng**, Siyang Liu, Minlie Huang. *PsyQA: A Chinese Dataset for Generating Long Counseling Text for Mental Health Support*. ACL 2021 Findings.
15. **Chujie Zheng**, Yunbo Cao, Daxin Jiang, Minlie Huang. *Difference-aware Knowledge Selection for Knowledge-grounded Conversation Generation*. EMNLP 2020 Findings.
16. Hao Zhou*, **Chujie Zheng***, Kaili Huang, Minlie Huang, Xiaoyan Zhu. *KdConv: A Chinese Multi-domain Dialogue Dataset Towards Multi-turn Knowledge-driven Conversation*. ACL 2020.
17. **Chujie Zheng**, Minlie Huang, Aixin Sun. *ChID: A Large-scale Chinese IDiom Dataset for Cloze Test*. ACL 2019.

Selected Awards and Honors

- Schlumberger Scholarship, Tsinghua University 2023
- Comprehensive Merit Scholarship, Tsinghua University 2022
- Outstanding Undergraduate, Tsinghua University 2020
- China National Scholarship (Top 2/100) 2019
- Comprehensive Merit Scholarship, Tsinghua University 2018

Academic Services

- **Area Chair:** ACL (24), ACL Rolling Review (24)
- **Conference Reviewer:** ACL (22/23), EMNLP (21/22), NAACL (24), EACL (23), ACL Rolling Review (21/22/23), CogSci (24), AAAI (22/23)
- **Journal Reviewer:** IEEE Transactions on Computational Social Systems (24), ACM Transactions on the Web (22), ACM Transactions on Intelligent Systems and Technology (22), Knowledge-Based Systems (21)