

Chujie Zheng 郑楚杰

Contact Information

Address(es)	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** ([preprint](#), [ICLR 2024 Spotlight](#), [ICML 2024](#)). My ultimate research goal is to **advance and oversee AI systems in a scalable way (with minimal human intervention) and ensure they operate transparently and responsibly**.
 - Previously, I conducted 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 **1000+**, h-index **15**, i10-index **18**.

Selected Papers

-
1. **Weak-to-Strong Extrapolation Expedites Alignment**
Chujie Zheng, Ziqi Wang, Heng Ji, Minlie Huang, Nanyun Peng
arXiv:2404.16792 (10K+ downloads on 🤗 HuggingFace in 2 weeks)
[\[paper\]](#) [\[repo\]](#) [🤗 HuggingFace](#)
 2. **On Prompt-Driven Safeguarding for Large Language Models**
Chujie Zheng, Fan Yin, Hao Zhou, Fandong Meng, Jie Zhou, Kai-Wei Chang, Minlie Huang, Nanyun Peng
ICML 2024 || Secure and Trustworthy LLM Workshop @ ICLR 2024 (Oral: 5%)
[\[paper\]](#) [\[repo\]](#)
 3. **Large Language Models Are Not Robust Multiple Choice Selectors**
Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, Minlie Huang
ICLR 2024 (Spotlight: 5%)

[\[paper\]](#) [\[repo\]](#)

4. **Click: Controllable Text Generation with Sequence Likelihood Contrastive Learning**

Chujie Zheng, Pei Ke, Zheng Zhang, Minlie Huang

ACL 2023 Findings

[\[paper\]](#) [\[repo\]](#)

5. **Towards Emotional Support Dialog Systems**

Siyang Liu*, **Chujie Zheng***, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, Minlie Huang (*: Equal contribution)

ACL 2021 (Oral)

[\[paper\]](#) [\[repo\]](#)

Other Papers

6. **AugESC: Dialogue Augmentation with Large Language Models for Emotional Support Conversation**

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, Minlie Huang

ACL 2023 Findings

7. **CASE: Aligning Coarse-to-Fine Cognition and Affection for Empathetic Response Generation**

Jinfeng Zhou*, **Chujie Zheng***, Bo Wang, Zheng Zhang, Minlie Huang

ACL 2023 (Oral)

8. **EVA2.0: Investigating Open-domain Chinese Dialogue Systems with Large-scale Pre-training**

Yuxian Gu*, Jiaxin Wen*, Hao Sun*, Yi Song, Pei Ke, **Chujie Zheng**, Zheng Zhang, Jianzhu Yao, Lei Liu, Xiaoyan Zhu, Minlie Huang

Machine Intelligence Research 2023

9. **CDConv: A Benchmark for Contradiction Detection in Chinese Conversations**

Chujie Zheng*, Jinfeng Zhou*, Yinhe Zheng, Libiao Peng, Zhen Guo, Wenquan Wu, Zhengyu Niu, Hua Wu, Minlie Huang

EMNLP 2022 (Oral)

10. **COLD: A Benchmark for Chinese Offensive Language Detection**

Jiawen Deng*, Jingyan Zhou*, Hao Sun, **Chujie Zheng**, Fei Mi, Helen Meng, Minlie Huang

EMNLP 2022 (Oral)

11. **On the Safety of Conversational Models: Taxonomy, Dataset, and Benchmark**

Hao Sun*, Guangxuan Xu*, Jiawen Deng, Jiale Cheng, **Chujie Zheng**, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, Minlie Huang

ACL 2022 Findings

12. **CEM: Commonsense-aware Empathetic Response Generation**

Sahand Sabour, **Chujie Zheng**, Minlie Huang

AAAI 2022 (Oral)

13. **Exploring Prompt-based Few-shot Learning for Grounded Dialog Generation**

Chujie Zheng, Minlie Huang

arXiv:2109.06513

14. **EVA: An Open-Domain Chinese Dialogue System with Large-Scale Generative Pre-Training**

Hao Zhou*, Pei Ke*, Zheng Zhang*, Yuxian Gu, Yinhe Zheng, **Chujie Zheng**, Yida Wang, Chen Henry Wu, Hao Sun, Xiacong Yang, Bosi Wen, Xiaoyan Zhu, Minlie Huang, Jie Tang
arXiv:2108.01547

15. **CoMAE: A Multi-factor Hierarchical Framework for Empathetic Response Generation**
Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, Minlie Huang
ACL 2021 Findings
16. **PsyQA: A Chinese Dataset for Generating Long Counseling Text for Mental Health Support**
Hao Sun*, Zhenru Lin*, **Chujie Zheng**, Siyang Liu, Minlie Huang
ACL 2021 Findings
17. **Difference-aware Knowledge Selection for Knowledge-grounded Conversation Generation**
Chujie Zheng, Yunbo Cao, Daxin Jiang, Minlie Huang
EMNLP 2020 Findings
18. **KdConv: A Chinese Multi-domain Dialogue Dataset Towards Multi-turn Knowledge-driven Conversation**
Hao Zhou*, **Chujie Zheng***, Kaili Huang, Minlie Huang, Xiaoyan Zhu
ACL 2020
19. **ChID: A Large-scale Chinese IDiom Dataset for Cloze Test**
Chujie Zheng, Minlie Huang, Aixin Sun
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)
 - **Reviewer:** NeurIPS (24), ICML (24), ACL (22/23), EMNLP (21/22), NAACL (24), EACL (23), ACL Rolling Review (21/22/23), CogSci (24), AAAI (22/23)