

# Chujie Zheng 郑楚杰

## Contact Information

---

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

## 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](#), [ACL 2023](#)). My research goal is to **advance and oversee AI systems in a scalable way (with minimal human intervention) and ensure they work responsibly and transparently**.
- 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 \(EMNLP 2022\)](#).
- I maintain the [GitHub repository](#) of **chat templates for LLMs**, which has received **480+** stars.
- Google Scholar citations **1300+**, h-index **17**, i10-index **18**.

## Education

---

- **Ph.D. candidate** in Computer Science and Technology, Tsinghua University Sep 2020 – Present  
Advisor: [Minlie Huang](#)
- **Visiting Scholar**, UCLA Nov 2023 – Jun 2024  
Host: [Nanyun \(Violet\) Peng](#)
- **B.S.** in Foundational Mathematics and Physics, Tsinghua University Aug 2016 – Jun 2020

## Work Experiences

---

- Research Intern. Qwen Post-training Team, **Alibaba Cloud** Oct 2024 – Present
- Research Intern. AI Alignment Team, **01.AI** Jul 2024 – Oct 2024  
Improved **Yi**'s mathematical reasoning via reward model enhancement.  
**Yi-Lightning** ranks **#6** on Chatbot Arena and **#3** in **Math** Category (as of 10/14/2024).
- Research Intern. WeChat AI Group, **Tencent** Feb 2023 – May 2024
- Research Intern. General Dialogue Team, **Baidu** Feb 2022 – Jun 2022  
Improved **PLATO**'s dialogue consistency via building contradiction detectors.

## Selected Projects/Papers

---

1. **Weak-to-Strong Extrapolation Expedites Alignment**  
Chujie Zheng, Ziqi Wang, Heng Ji, Minlie Huang, Nanyun Peng  
*arXiv:2404.16792 (20+ citations by DeepMind, Oxford, etc.)*  
**70K+ downloads on 🤗 HuggingFace (10K+ 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 || SeT LLM Workshop @ ICLR 2024 (Oral: 5%) (40+ citations by Anthropic, MIT, etc.)*  
[\[paper\]](#) [\[repo\]](#)
3. **Chat Templates for 🤗 HuggingFace Large Language Models**  
Chujie Zheng  
*GitHub Repository (480+ stars)*  
[\[repo\]](#)
4. **Large Language Models Are Not Robust Multiple Choice Selectors**  
Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, Minlie Huang  
*ICLR 2024 (Spotlight: 5%) (110+ citations by Meta's LLaMA-3, DeepMind, etc.)*  
[\[paper\]](#) [\[repo\]](#)
5. **Click: Controllable Text Generation with Sequence Likelihood Contrastive Learning**  
Chujie Zheng, Pei Ke, Zheng Zhang, Minlie Huang  
*Findings of ACL 2023*  
**Early work of preference optimization. Completed in 2022**  
[\[paper\]](#) [\[repo\]](#)
6. **AugESC: Dialogue Augmentation with Large Language Models for Emotional Support Conversation**  
Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, Minlie Huang  
*Findings of ACL 2023*  
**Early work of LLM-based data synthesis. Completed in 2021**  
[\[paper\]](#) [\[repo\]](#)
7. **Towards Emotional Support Dialog Systems**  
Siyang Liu\*, Chujie Zheng\*, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, Minlie Huang (\*: Equal contribution)  
*ACL-IJCNLP 2021 (Oral) (220+ citations)*  
[\[paper\]](#) [\[repo\]](#)

## Other Papers

---

8. **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)*

9. **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*
10. **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)*
11. **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)*
12. **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  
*Findings of ACL 2022*
13. **CEM: Commonsense-aware Empathetic Response Generation**  
Sahand Sabour, Chujie Zheng, Minlie Huang  
*AAAI 2022 (Oral) (150+ citations)*
14. **Exploring Prompt-based Few-shot Learning for Grounded Dialog Generation**  
Chujie Zheng, Minlie Huang  
*arXiv:2109.06513*
15. **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, Xiaocong Yang, Bosi Wen, Xiaoyan Zhu, Minlie Huang, Jie Tang  
*arXiv:2108.01547*
16. **CoMAE: A Multi-factor Hierarchical Framework for Empathetic Response Generation**  
Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, Minlie Huang  
*Findings of ACL-IJCNLP 2021*
17. **PsyQA: A Chinese Dataset for Generating Long Counseling Text for Mental Health Support**  
Hao Sun\*, Zhenru Lin\*, Chujie Zheng, Siyang Liu, Minlie Huang  
*Findings of ACL-IJCNLP 2021*
18. **Difference-aware Knowledge Selection for Knowledge-grounded Conversation Generation**  
Chujie Zheng, Yunbo Cao, Daxin Jiang, Minlie Huang  
*Findings of EMNLP 2020*
19. **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 (110+ citations)*
20. **ChID: A Large-scale Chinese IDiom Dataset for Cloze Test**  
Chujie Zheng, Minlie Huang, Aixin Sun  
*ACL 2019 (80+ citations)*

## 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), EMNLP (24), NAACL (25), ACL Rolling Review (24)
- **Reviewer:** ICLR (25), NeurIPS (24), ICML (24), COLM (24), ACL (22/23), EMNLP (21/22), NAACL (24), EACL (23), ACL Rolling Review (21/22/23), CogSci (24), AAAI (22/23)