郑锐 (RuiZheng)

15156543297 rzheng20@fudan.edu.cn
PHD Student

EDUCATION

Fudan University Sep 2020 - Jun 2024

PHD in Computer Science

- GPA: 3.38/ 4.0 (Top 10%)
- Honors/Awards: Tencent Scholarship (2022)

University of Science and Technology of China Sep 2017 - Jun 2020

M.S. in Electrical Engineering

- GPA: 3.8 / 4.3 (Top 10%)
- National Scholarship (2019)

PUBLICATION

- **Rui Zheng**, Shihan Dou, Songyang Gao et al., Secrets of RLHF in Large Language Models Part I: PPO, arXiv:2307.04964.
- **Rui Zheng**, Wei Shen, Yuan Hua et al., Improving Generalization of Alignment with Human Preferences through Group Invariant Learning, arXiv:2310.11971.
- **Rui Zheng**, Zhiheng Xi et al., Characterizing the Impacts of Instances on Robustness, ACL findings 2023, CCF-A.
- **Rui Zheng**, Shihan Dou, Yuhao Zhou et al., Detecting Adversarial Samples through Sharpness of Loss Landscape, ACL findings 2023, CCF-A.
- **Rui Zheng**, Rong Bao et al., Robust Lottery Tickets for Pre-trained Language Models, ACL 2022, CCF-A.
- **Rui Zheng**, Rong Bao et al., PlugAT: A Plug and Play Module to Defend against Textual Adversarial Attack, COLING 2022, CCF-B.
- Rong Bao, **Rui Zheng** (Equal Contribution) et al., CASN:Class-Aware Score Network for Textual Adversarial Detection, ACL 2023 CCF-A.
- Zhiheng Xi, **Rui Zheng** (Equal Contribution) et al., Connectivity Patterns are Task Embeddings, ACL findings 2023, CCF-A.
- Ting Wu, **Rui Zheng** (Equal Contribution) et al., Modeling the Q-Diversity in a Min-max Play Game for Robust Optimization, ACL findings 2023, CCF-A.
- Wei Shen, **Rui Zheng** (Equal Contribution) et al., Loose lips sink ships: Mitigating Length Bias in Reinforcement Learning from Human Feedback, EMNLP findings 2023, CCF-B.
- Enyu Zhou, **Rui Zheng** (Equal Contribution) et al., RealBehavior: A Framework for Faithfully Characterizing Foundation Models' Human-like Behavior Mechanisms, EMNLP findings 2023, CCF-B.
- Qin Liu, **Rui Zheng** (Equal Contribution) et al., Flooding-X: Improving BERT's Resistance to Adversarial Attacks via Loss-Restricted Fine-Tuning, ACL 2022, CCF-A.
- Zhiheng Xi, **Rui Zheng** (Equal Contribution) et al., Efficient Adversarial Training with Robust Early-Bird Tickets, EMNLP 2022, CCF-B.
- Shihan Dou, **Rui Zheng** (Equal Contribution) et al., Decorrelate Irrelevant, Purify Relevant: Overcome Textual Spurious Correlations from a Feature Perspective, COLING 2022, CCF-B.

OPEN SOURCE PROJECT

MOSS-RLHF Aug 2023 - Present

Leader

https://github.com/OpenLMLab/MOSS-RLHF

TextFlint Nov 2020 - May 2021

Major Contributor

https://github.com/textflint/textflint

INTERNSHIP EXPERIENCE

ByteDance AI Lab Mar 2023 - Present

Research on RLHF Pipeline

IFLYTEK Jun 2020 - Aug 2020

Research on Speech Enhancement