Zhexin Zhang

+86 18810058042 | zx-zhang22@mails.tsinghua.edu.cn
Webpage: https://nonstopfor.github.io/

EDUCATION

Tsinghua University Sep 2022 - Jun 2027
Ph.D. Student, Department of Computer Science and Technology
- CoAI Group, advised by Prof. Minlie Huang.

Tsinghua University Sep 2018 - Jun 2022
B.Eng., Department of Computer Science and Technology
- GPA: 3.90/4.00 (Top 6%)

PUBLICATIONS

AISafetyLab: A Comprehensive Framework for AI Safety Evaluation and Improvement Sep 2024 - Jan 2025
- Will be submitted to ACL Demo. First author.
- A comprehensive framework and toolkit covering various safety attack, defense and evaluation methods.

Agent-SafetyBench: Evaluating the Safety of LLM Agents Aug 2024 - Dec 2024
- Submitted to ACL ARR. First author.
- A comprehensive agent safety evaluation benchmark which incorporates 349 interaction environments.

Safe Unlearning: A Surprisingly Effective and Generalizable Solution to Defend Against Jailbreak Attacks May 2024 - Jan 2025
- Will be submitted to ICML. First author.
- Identifying and analyzing the surprising generalization ability of unlearning to defend against jailbreak attacks.

ShieldLM: Empowering LLMs as Aligned, Customizable and Explainable Safety Detectors Nov 2023 - Jun 2024
- Findings of EMNLP 2024. First author.
- The first specialized safety detector that supports fine-grained safety rules. It has been downloaded for more than 40K times on Huggingface.

Defending Large Language Models Against Jailbreaking Attacks Through Goal Prioritization Oct 2023 - Feb 2024
- ACL 2024. First author. Has been cited 80 times on Semantic Scholar.
- Implementing goal (safety and helpfulness) prioritization to defend against jailbreak attacks.

SafetyBench: Evaluating the Safety of Large Language Models Jun 2023 - Oct 2023
- ACL 2024. First author. Has been cited 116 times on Google Scholar.
- The first comprehensive LLM safety evaluation benchmark based on multiple-choice questions. It has been integrated into SuperBench and OpenCompass. We have received more than 4K test requests.

InstructSafety: A Unified Framework for Building Multidimensional and Explainable Safety Detector through Instruction Tuning Jan 2023 - May 2023
- Findings of EMNLP 2023. First author.
- Proposea unified safety detection framework based on instruction tuning.

Ethicist: Targeted Training Data Extraction Through Loss Smoothed Soft Prompting and Calibrated Confidence Estimation Aug 2022 - Nov 2022
- ACL 2023 (Oral). First author.
- Achieves the SOTA performance on the targeted training data extraction task.

Constructing Highly Inductive Contexts for Dialogue Safety through Controllable Reverse Generation Jan 2022 - Jun 2022
- Findings of EMNLP 2022. First author.
- Automatically construct contexts that can induce unsafe responses through reserve generation.

HONORS & AWARDS

Top 40 Global Finalist of the Baidu Scholarship 2024
National Scholarship, Dept. CST, Tsinghua University 2024
Third place, Global Challenge for Safe and Secure LLMs 2024
Samsung Scholarship, Dept. CST, Tsinghua University 2023
Excellent Graduate, Tsinghua University 2022
Outstanding Graduate, Dept. CST, Tsinghua University 2022
Third place (3/3665), WeChat Big Data Challenge Finals 2022
Sixth place (6/1000+), Global AI Innovation Contest Finals 2022
Second place (2/5000+), Global AI Innovation Contest Finals 2021
Academic Excellence Scholarship, Dept. CST, Tsinghua University 2020,2021
Meritorious Winner (<10%), The Mathematics Contest in Modeling 2020