Hongpeng Lin

https://hopelin99.github.io | Google Scholar | hopelin@ruc.edu.cn | (+86) 187-597-90891

Research Interests

My research interests are in Natural Language Processing and Multi-modal Understanding, with the hope of making machines perceive, understand, and express like humans. Currently, my research focus is committed to adversarial robustness (jailbreaking LLMs and mitigation), long video understanding (natural language query localization), and creative multi-modal text generation (humor and association).

Education

Renmin University of China, Gaoling School of Artificial Intelligence Sep. 2021 – Jun. 2024
- Master of Science in Artificial Intelligence. *Advisor:* Ruihua Song

Xidian University Sep. 2017 – Jun. 2021
- Bachelor of Engineering in Software Engineering.

Experience

Research Intern, Stanford University Sep. 2023 - Jan. 2024
- Social Influence Jailbreak Program. *Advisor:* Weiyan Shi, Yi Zeng

Intern, Xianyuan Technology Jun. 2023 - Aug.2023
- Studies of LLMs and MLLMs.

Intern, Kuaishou Technology Nov. 2021 - Jun. 2022
- WenLan project of large-scale multi-modal pre-training.

Research Intern, Beijing Academy of Artificial Intelligence (BAAI) Oct. 2021 - Jan. 2022
- Multi-modal data crawling and pre-processing.

Research Intern, Tencent AI Lab Jul. 2021 - Aug. 2021
- Development of a multi-modal dialogue WeChat mini-program.

Publications

[*: Equal Contribution]

(1) How Johnny Can Persuade LLMs to Jailbreak Them: Rethinking Persuasion to Challenge AI Safety by Humanizing LLMs
Yi Zeng*, **Hongpeng Lin***, Jingwen Zhang, Diyi Yang, Ruoxi Jia, Weiyan Shi
arXiv Preprint 2024.

(2) TikTalk: A Video-Based Dialogue Dataset for Multi-Modal Chitchat in Real World
Hongpeng Lin*, Ludan Ruan*, Wenke Xia*, Peiyu Liu, Jingyuan Wen, Yixin Xu, Di Hu, Ruihua Song, Wayne Xin Zhao, Qin Jin, Zhiwu Lu.
ACM MM 2023 (Oral).

(3) An Analysis of Collection Methods and Simulation Algorithms for Chinese Speech Errors
Hao Pu, **Hongpeng Lin**, Ruihua Song, Mei Yan.
CLSW 2022.

Awards

Graduate Student First Class Academic Scholarship, 2023

Skills

Technical Skills: Proficient in Python and Pytorch, and has practical experience in finetuning (M)LLMs.

Characteristic: Reliable and responsible, with a strong sense of teamwork and a passion for research.

Academic Services

Conference Reviewer: EMNLP, ACM MM, CVPR, ICML