KEXIN HUANG

huangkx19@gmail.com ◇ Google Scholar ◇ GitHub ◇ Homepage

Last Update: July 5, 2024

EDUCATION

Fudan University *Shanghai, China*
MSc in Computer Science and Technology *Sep 2024 - Present*
Advisor: Prof. Xipeng QIU

Fudan University *Shanghai, China*
BSc in Computer Science and Technology *Sep 2019 - Jun 2023*
Advisor: Prof. Yang CHEN

WORKING EXPERIENCE

Research Assistant, AI Governance, Shanghai AI Laboratory *Shanghai, China*
Advisor: Dr. Yan TENG *Dec 2022 - Jul 2024*
Internship (*Dec 2022 - Jun 2023*), Employment (*Jul 2023 - Jul 2024*)

RESEARCH INTEREST

Large Language Model / Alignment / Natural Language Processing

PUBLICATIONS

MLLMGuard: A Multi-dimensional Safety Evaluation Suite for Multimodal Large Language Models (*arXiv*) *Jun 2024*
Tianle Gu, Zeyang Zhou, **Kexin Huang**, Dandan Liang, Yixu Wang, Haiquan Zhao, Yuanqi Yao, Xingge Qiao, Keqing Wang, Yujiu Yang, Yan Teng, Yu Qiao, Yingchun Wang

ESC-Eval: Evaluating Emotion Support Conversations in Large Language Models (*arXiv*) *Jun 2024*
Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, **Kexin Huang**, Tianle Gu, Yixu Wang, Dandan Liang, Zhixu Li, Yan Teng, Yanghua Xiao, Yingchun Wang

From Pixels to Principles: A Decade of Progress and Landscape in Trustworthy Computer Vision (*Science and Engineering Ethics 30 (3), 26*) *Jun 2024*
Kexin Huang, Yan Teng, Yang Chen, Yingchun Wang

From GPT-4 to Gemini and Beyond: Assessing the Landscape of MLLMs on Generalizability, Trustworthiness and Causality through Four Modalities (*arXiv*) *Jan 2024*
Chaochao Lu, Chen Qian, **Kexin Huang**, et al.

Flames: Benchmarking Value Alignment of LLMs in Chinese (*NAACL 2024*) *Nov 2023*
Kexin Huang*, Xiangyang Liu*, Qianyu Guo*, Tianxiang Sun, Jiawei Sun, Yaru Wang, Zeyang Zhou, Yixu Wang, Yan Teng, Xipeng Qiu, Yingchun Wang, Dahua Lin (* denotes equal contributions.)

Fake Alignment: Are LLMs Really Aligned Well? (*NAACL 2024*) *Nov 2023*
Yixu Wang, Yan Teng, **Kexin Huang**, Chengqi Lyu, Songyang Zhang, Wenwei Zhang, Xingjun Ma, Yu-Gang Jiang, Yu Qiao, Yingchun Wang

SKILLS

Computer	Python, Pytorch, C, C++, HTML, SQL, MATLAB, Photoshop, Office
Language	Chinese (native), English (fluent - TOEFL iBT 108), Hokkien (conversational)
Interest	Travelling, Painting, Photography, Tennis, Badminton