Jen-Tse (Jay) Huang

XBzDTAQAAAAJ 0000-0003-3446-0083 2161306685 317/7026 08a169200 penguinnnnn
✉ jthuang@cse.cuhk.edu.hk https://penguinnnnn.github.io/ Sha Tin, Hong Kong

Education

Ph.D. in Computer Science, The Chinese University of Hong Kong	*Aug. 2020 - Jan. 2025*
B.Sc. in Computer Science, Peking University	*Sep. 2015 - Jul. 2019*

Experience

Visiting Student, University of Southern California, Los Angeles, CA	*Jul. 2024 - Dec. 2024*
Research Intern, Tencent AI Lab, Shenzhen	*Mar. 2022 - Oct. 2023*
Research Assistant, The Chinese University of Hong Kong, Hong Kong	*Feb. 2020 - Jul. 2020*
Research Intern, SenseTime Research, Beijing	*Feb. 2018 - Jun. 2019*

Publications

* equal contribution ✉ corresponding author

[17] **Jen-tse Huang**, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao ✉, Zhaopeng Tu, Michael R. Lyu, 2024. Apathetic or Empathetic? Evaluating LLMs' Emotional Alignments with Humans. In *Advances in Neural Information Processing Systems 37.* (NeurIPS'24)

[16] **Jen-tse Huang**, Wenxiang Jiao, Man Ho Lam, Eric John Li, Wenxuan Wang ✉, Michael R. Lyu, 2024. On the Reliability of Psychological Scales on Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing.* (EMNLP Main'24)

[15] Ziyi Liu *, Abhishek Anand *, Pei Zhou, **Jen-tse Huang**, Jieyu Zhao, 2024. InterIntent: Investigating Social Intelligence of LLMs via Intention Understanding in an Interactive Game Context. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing.* (EMNLP Main'24)

[14] Yuxuan Wan *, Wenxuan Wang *, Wenxiang Jiao, Yiliu Yang, Youliang Yuan, **Jen-tse Huang**, Pinjia He, Michael R. Lyu, 2024. LogicAsker: Evaluating and Improving the Logical Reasoning Ability of Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing.* (EMNLP Main'24)

[13] Wenxuan Wang, Haonan Bai, Yuxuan Wan, **Jen-tse Huang** ✉, Youliang Yuan, Haoyi Qiu, Nanyun Peng, Michael R. Lyu, 2024. New Job, New Gender? Measuring the Social Bias in Image Generation Models. In *Proceedings of the 32nd ACM Multimedia Conference.* (ACMMM'24)
[Oral Presentation (174/4385, 3.97%)]

[12] Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, **Jen-tse Huang**, Zhaopeng Tu ✉, Michael R. Lyu, 2024. Not All Countries Celebrate Thanksgiving: On the Cultural Dominance in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* pp. 6349–6384. (ACL Main'24)

[11] Xintao Wang, Yunze Xiao, **Jen-tse Huang**, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Cheng Li, Jiangjie Chen, Wei Wang, Yanghua Xiao ✉, 2024. InCharacter: Evaluating Personality Fidelity in Role-Playing Agents through Psychological Interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* pp. 1840–1873. (ACL Main'24)

[10] Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, **Jen-tse Huang** ✉, Wenxiang Jiao, Michael R. Lyu, 2024. All Languages Matter: On the Multilingual Safety of Large Language Models. In *Findings of the Association for Computational Linguistics ACL 2024.* pp. 5865–5877. (ACL Findings'24)

[9] **Jen-tse Huang**, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao ✉, Zhaopeng Tu, Michael R. Lyu, 2024. On the Humanity of Conversational AI: Evaluating the Psychological Portrayal of LLMs. In *Proceedings of the Twelfth International Conference on Learning*

Representations. pp. 1-24. (ICLR'24)
[Oral Presentation (86/7404, 1.16%)]

[8] Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, **Jen-tse Huang**, Pinjia He ✉, Shuming Shi, Zhaopeng Tu, 2024. GPT-4 Is Too Smart To Be Safe: Stealthy Chat with LLMs via Cipher. In *Proceedings of the Twelfth International Conference on Learning Representations*. pp. 1-21. (ICLR'24)

[7] Wenxiang Jiao ✉, **Jen-tse Huang**, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, Zhaopeng Tu, 2023. ParroT: Translating During Chat Using Large Language Models tuned with Human Translation and Feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. pp. 15009-15020. (EMNLP Findings'23)

[6] Wenxuan Wang, Jingyuan Huang, **Jen-tse Huang**, Chang Chen, Jiazhen Gu ✉, Pinjia He, Michael R. Lyu, 2023. An Image is Worth a Thousand Toxic Words: A Metamorphic Testing Framework for Content Moderation Software. In *2023 38th IEEE/ACM International Conference on Automated Software Engineering*. pp. 1339-1351. (ASE'23)

[5] Jianping Zhang, **Jen-tse Huang**, Wenxuan Wang, Yichen Li, Weibin Wu ✉, Xiaosen Wang, Yuxin Su, Michael Lyu, 2023. Improving the Transferability of Adversarial Samples by Path-Augmented Method. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8173-8182. (CVPR'23)

[4] Wenxuan Wang, **Jen-tse Huang**, Weibin Wu, Jianping Zhang, Yizhan Huang, Shuqing Li, Pinjia He ✉, Michael R. Lyu, 2023. MTTM: Metamorphic Testing for Textual Content Moderation Software. In *2023 IEEE/ACM 45th International Conference on Software Engineering*. pp. 2387-2399. (ICSE'23)

[3] Wenxiang Jiao, Zhaopeng Tu, Jiarui Li, Wenxuan Wang, **Jen-tse Huang**, Shuming Shi, 2022. Tencent's Multilingual Machine Translation System for WMT22 Large-Scale African Languages. In *Proceedings of the Seventh Conference on Machine Translation*. pp. 1049–1056. (WMT'22)

[2] **Jen-tse Huang**, Jianping Zhang, Wenxuan Wang, Pinjia He ✉, Yuxin Su, Michael R. Lyu, 2022. AEON: A Method for Automatic Evaluation of NLP Test Cases. In *Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis*. pp. 202-214. (ISSTA'22)

[1] Jianping Zhang, Weibin Wu ✉, **Jen-tse Huang**, Yizhan Huang, Wenxuan Wang, Yuxin Su, Michael R. Lyu, 2022. Improving Adversarial Transferability via Neuron Attribution-Based Attacks. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 14973-14982. (CVPR'22)

Preprints

[12] **Jen-tse Huang**, Jiaxu Zhou, Tailin Jin, Xuhui Zhou, Zixi Chen, Wenxuan Wang ✉, Youliang Yuan, Maarten Sap, Michael R. Lyu, 2024. On the Resilience of Multi-Agent Systems with Malicious Agents. *arXiv Preprint: 2408.00989*.

[11] **Jen-tse Huang**, Eric John Li, Man Ho Lam, Tian Liang, Wenxuan Wang, Youliang Yuan, Wenxiang Jiao ✉, Xing Wang, Zhaopeng Tu, Michael R. Lyu, 2024. How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments. *arXiv Preprint: 2403.11807*.

[10] Wenxuan Wang, Kuiyi Gao, Zihan Jia, Youliang Yuan, **Jen-tse Huang**, Qiuzhi Liu, Shuai Wang, Wenxiang Jiao ✉, Zhaopeng Tu ✉, 2024. Chain-of-Jailbreak Attack for Image Generation Models via Editing Step by Step. *arXiv Preprint: 2410.03869*.

[9] Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, **Jen-tse Huang**, Jiahao Xu, Tian Liang, Pinjia He ✉, Zhaopeng Tu, 2024. Refuse Whenever You Feel Unsafe: Improving Safety in LLMs via Decoupled Refusal Training. *arXiv Preprint: 2407.09121*.

[8] Xiaoyuan Liu, Wenxuan Wang, Youliang Yuan, **Jen-tse Huang**, Qiuzhi Liu, Pinjia He ✉, Zhaopeng Tu, 2024. Insight Over Sight? Exploring the Vision-Knowledge Conflicts in Multimodal LLMs. *arXiv Preprint: 2410.08145*.

[7] Jen-Yuan Huang, Haofan Wang, Qixun Wang, Xu Bai, Hao Ai, Peng Xing, **Jen-tse Huang**, 2024. InstantIR: Blind Image Restoration with Instant Generative Reference. *arXiv Preprint: 2410.06551*.

[6] Wenxuan Wang, Juluan Shi, Chaozheng Wang, Cheryl Lee, Youliang Yuan, **Jen-tse Huang** ✉, Michael R. Lyu, 2024. Learning to Ask: When LLMs Meet Unclear Instruction. *arXiv Preprint: 2409.00557*.

[5] Wenxuan Wang *, Juluan Shi *, Zhaopeng Tu, Youliang Yuan, **Jen-tse Huang**, Wenxiang Jiao, Michael R. Lyu, 2024. The Earth is Flat? Unveiling Factual Errors in Large Language Models. *arXiv Preprint: 2401.00761*.

[4] Cheryl Lee, Chunqiu Steven Xia, **Jen-tse Huang**, Zhouruixin Zhu, Lingming Zhang, Michael R. Lyu, 2024. A Unified Debugging Approach via LLM-Based Multi-Agent Synergy. *arXiv Preprint: 2404.17153*.

[3] Man Tik Ng *, Hui Tung Tse *, **Jen-tse Huang** ✉, Jingjing Li, Wenxuan Wang, Michael R. Lyu, 2024. How Well Can LLMs Echo Us? Evaluating AI Chatbots' Role-Play Ability with ECHO. *arXiv Preprint: 2404.13957*.

[2] Tian Liang, Zhiwei He, **Jen-tse Huang**, Wenxuan Wang, Wenxiang Jiao ✉, Rui Wang, Yujiu Yang ✉, Zhaopeng Tu, Shuming Shi, Xing Wang ✉, 2023. Leveraging Word Guessing Games to Assess the Intelligence of Large Language Models. *arXiv Preprints: 2310.20499*.

[1] Wenxiang Jiao ✉, Wenxuan Wang, **Jen-tse Huang**, Xing Wang, Shuming Shi, Zhaopeng Tu, 2023. Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine. *arXiv Preprints: 2301.08745*.

Service

- Conference Reviewer: ICLR'25; NeurIPS'24; CVPR'24; ACL'23; EMNLP'23,24
- Teaching Assistant: Software Engineering (CSCI3100) in CUHK *2021, 2022*
- Teaching Assistant: Discrete Math for Engineers (ENGG2440A) in CUHK *2020*