

# JEN-TSE (JAY) HUANG

✉ XBzDTAQAAAAJ    ID 0000-0003-3446-0083    ↗ 2161306685    ↙ 317/7026    IN 08a169200    GitHub penguinnnnn  
✉ jhuan236@jh.edu    🌐 <https://penguinnnnn.github.io/>    📍 Baltimore, Maryland

## EXPERIENCE

Postdoctoral Researcher, Johns Hopkins University, Baltimore, MD	Feb. 2025 - Present
Visiting Researcher, University of Southern California, Los Angeles, CA	Jul. 2024 - Dec. 2024
Research Intern, Tencent AI Lab, Shenzhen	Feb. 2022 - Jul. 2024
Research Assistant, Chinese University of Hong Kong, Hong Kong	Feb. 2020 - Jul. 2020
Research Intern, SenseTime Research, Beijing	Feb. 2018 - Jul. 2019

## EDUCATION

Ph.D. in Computer Science and Engineering, Chinese University of Hong Kong	Aug. 2020 - Dec. 2024
B.Sc. in Computer Science, Peking University	Sep. 2015 - Jul. 2019

## AWARD AND HONOR

• KAUST Rising Stars in AI	2026
• Outstanding Reviewer	EMNLP 2024
• Top Reviewer [1304/15160, 8.6%]	NeurIPS 2024
• CUHK PhD IMPAC Award, funded with HK\$50,000	2024

## TEACHING

• Teaching Assistant: Software Engineering (CSCI3100) at CUHK	2021, 2022
• Teaching Assistant: Discrete Math for Engineers (ENGG2440A) at CUHK	2020

## SERVICE

• Conference Area Chair: <b>2026</b> : ACL (ARR Jan'26), EACL (ARR Oct'25).	
• Conference Reviewer: <b>2026</b> : ICML, CVPR, ICLR; <b>2025</b> : IJCNLP-AACL (ARR Jul'25), EMNLP (ARR May'25), NeurIPS, NLPCC, ICCV, ACL (ARR Feb'25), ICML, NAACL (ARR Oct'24), CVPR, ICLR; <b>2024</b> : ARR Dec'24, EMNLP (ARR Jun'24), NeurIPS, CVPR; <b>2023</b> : EMNLP, ACL.	
• Journal Reviewer: Nature Human Behavior; Transactions on Machine Learning Research;	
• Organizer: 1st PersonaLLM @NeurIPS 2025, NENLP'25;	

## CONFERENCE PAPERS

\* equal contribution    ✉ corresponding author

- [32] Youliang Yuan, Wenxiang Jiao, Yuejin Xie, Chihao Shen, Menghan Tian, Wenxuan Wang, **Jen-tse Huang**, Pinjia He, 2025. Towards Evaluating Proactive Risk Awareness of Multimodal Language Models. *In Advances in Neural Information Processing Systems 38*. (NeurIPS'25)
- [31] Man Ho Lam, Chaozheng Wang ✉, **Jen-tse Huang**, Michael R. Lyu, 2025. CodeCrash: Exposing LLM Fragility to Misleading Natural Language in Code Reasoning. *In Advances in Neural Information Processing Systems 38*. (NeurIPS'25)
- [30] **Jen-tse Huang** \*, Jiantong Qin \*, Jianping Zhang, Youliang Yuan, Wenxuan Wang ✉, Jieyu Zhao ✉, 2025. VisBias: Measuring Explicit and Implicit Social Biases in Vision Language Models. *In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17981-18004. (EMNLP Main'25)

- [29] Jingyuan Huang \*, Jen-tse Huang \*, Ziyi Liu, Xiaoyuan Liu, Wenxuan Wang✉, Jieyu Zhao✉, 2025. AI Sees Your Location—But With A Bias Toward The Wealthy World. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 18030-18050. (EMNLP Main'25)
- [28] Jen-tse Huang, Yuhang Yan \*, Linqi Liu \*, Yixin Wan, Wenxuan Wang✉, Kai-Wei Chang, Michael R. Lyu, 2025. Where Fact Ends and Fairness Begins: Redefining AI Bias Evaluation through Cognitive Biases. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 10974-10993. (EMNLP Findings'25)
- [27] Wenxuan Wang, Juluan Shi \*, Zixuan Ling \*, Yuk-Kit Chan \*, Chaozheng Wang, Cheryl Lee, Youliang Yuan, Jen-tse Huang✉, Wenxiang Jiao✉, Michael R. Lyu, 2025. Learning to Ask: When LLM Agents Meet Unclear Instruction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21784-21795. (EMNLP Main'25)
- [26] Cheryl Lee✉, Chunqiu Steven Xia, Longji Yang, Jen-tse Huang, Zhouruixin Zhu, Lingming Zhang, Michael R. Lyu, 2025. UniDebugger: Hierarchical Multi-Agent Framework for Unified Software Debugging. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 18248-18277. (EMNLP Main'25)
- [25] Wenxuan Wang, Xiaoyuan Liu, Kuiyi Gao, Jen-tse Huang, Youliang Yuan, Pinjia He, Shuai Wang, Zhaopeng Tu✉, 2025. Can't See the Forest for the Trees: Benchmarking Multimodal Safety Awareness for Multimodal LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16993-17006. (ACL Main'25)
- [24] Xiaoyuan Liu, Wenxuan Wang, Youliang Yuan, Jen-tse Huang, Qiuwei Liu, Pinjia He✉, Zhaopeng Tu, 2025. Insight Over Sight: Exploring the Vision-Knowledge Conflicts in Multimodal LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17825-17846. (ACL Main'25)
- [23] Wenxuan Wang, Kuiyi Gao, Youliang Yuan, Jen-tse Huang, Qiuwei Liu, Shuai Wang, Wenxiang Jiao✉, Zhaopeng Tu✉, 2025. Chain-of-Jailbreak Attack for Image Generation Models via Step by Step Editing. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10940-10957. (ACL Findings'25)
- [22] Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Jiahao Xu, Tian Liang, Pinjia He✉, Zhaopeng Tu, 2025. Refuse Whenever You Feel Unsafe: Improving Safety in LLMs via Decoupled Refusal Training. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3149-3167. (ACL Main'25)
- [21] Jen-tse Huang, Jiaxu Zhou, Tailin Jin, Xuhui Zhou, Zixi Chen, Wenxuan Wang✉, Youliang Yuan, Michael R. Lyu, Maarten Sap✉, 2025. On the Resilience of LLM-Based Multi-Agent Collaboration with Faulty Agents. In *Proceedings of the 42nd International Conference on Machine Learning, PMLR vol. 267*, pp. 26202-26226. (ICML'25)
- [20] Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, Yanghua Xiao✉, 2025. CoSER: Coordinating LLM-Based Persona Simulation of Established Roles. In *Proceedings of the 42nd International Conference on Machine Learning, PMLR vol. 267*, pp. 64822-64858. (ICML'25)
- [19] Xuhui Zhou \*, Zhe Su \*, Sophie Feng, Jiaxu Zhou, Jen-tse Huang, Hsien-Te Kao, Spencer Lynch, Svitlana Volkova, Tongshuang Wu, Anita Woolley, Hao Zhu, Maarten Sap, 2025. SOTONIA-S4: A User-Friendly System for Flexible, Customizable, and Large-Scale Social Simulation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pp. 350-360. (NAACL Demo'25)
- [18] Jen-tse Huang, Eric John Li, Man Ho Lam, Tian Liang, Wenxuan Wang✉, Youliang Yuan, Wenxiang Jiao✉, Xing Wang, Zhaopeng Tu, Michael R. Lyu, 2025. Competing Large Language Models in Multi-Agent Gaming Environments. In *the Thirteenth International Conference on Learning Representations*. (ICLR'25)

- [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*, pp. 97053-97087. (NeurIPS'24)
- [16] 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*, pp. 6718-6746. (EMNLP Main'24)
- [15] Yuxuan Wan \*, Wenxuan Wang \*, Yiliu Yang, Youliang Yuan, **Jen-tse Huang**, Pinjia He, Wenxiang Jiao✉, 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*, pp. 2124-2155. (EMNLP Main'24)
- [14] **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*, pp. 6152-6173. (EMNLP Main'24)
- [13] Wenxuan Wang, Haonan Bai, **Jen-tse Huang**✉, Yuxuan Wan, 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 International Conference on Multimedia*, pp. 3781-3789. (ACMMM'24)  
[Oral 174/4385 3.97%]
- [12] Xintao Wang, Yunze Xiao, **Jen-tse Huang**, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, 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)
- [11] 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 LLMs. *In Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5865-5877. (ACL Findings'24)
- [10] 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)
- [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 the Twelfth International Conference on Learning Representations*. (ICLR'24)  
[Oral 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 the Twelfth International Conference on Learning Representations*. (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 Proceedings of the 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 R. Lyu, 2023. Improving the Transferability of Adversarial Samples by Path-Augmented

Method. In *Proceedings of the 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 *Proceedings of the 45th IEEE/ACM 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)  
[1st Place in this Track]
- [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 *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14993-15002. (CVPR'22)

#### JOURNAL PAPERS

---

- [1] Wenxuan Wang, Wenxiang Jiao, **Jen-tse Huang**, Zhaopeng Tu✉, Michael R. Lyu, 2025. On the Shortcut Learning in Multilingual Neural Machine Translation. *Neurocomputing*, vol. 615, no. 128833. (Neurocomputing'25)

#### PREPRINT PAPERS

---

- [26] Xintao Wang \*, Jian Yang \*, Weiyuan Li, Rui Xie, **Jen-tse Huang**, Jun Gao, Shuai Huang, Yueping Kang, Liyuan Gou, Hongwei Feng, Yanghua Xiao✉, 2026. HumanLLM: Benchmarking and Reinforcing LLM Anthropomorphism via Human Cognitive Patterns. *arXiv Preprint: 2601.10198*
- [25] **Jen-tse Huang**✉\*, Jiantong Qin \*, Xueli Qiu, Sharon Levy, Michelle R. Kaufman, Mark Dredze, 2026. Knowing But Not Doing: Convergent Morality and Divergent Action in LLMs. *arXiv Preprint: 2601.07972*
- [24] **Jen-tse Huang** \*, Chang Chen \*, Shiyang Lai, Wenxuan Wang✉, Michelle R. Kaufman, Mark Dredze, 2026. Probing Multimodal Large Language Models on Cognitive Biases in Chinese Short-Video Misinformation. *arXiv Preprint: 2601.06600*
- [23] Boyuan Chen, Sitong Fang, Jiaming Ji, Yanxu Zhu, Pengcheng Wen, Jinzhou Wu, Yingshui Tan, Boren Zheng, Mengying Yuan, Wenqi Chen, Donghai Hong, Alex Qiu, Xin Chen, Jiayi Zhou, Kaile Wang, Juntao Dai, Borong Zhang, Tianzhuo Yang, Saad Siddiqui, Isabella Duan, Yawen Duan, Brian Tse, **Jen-tse Huang**, Kun Wang, Baihui Zheng, Jiaheng Liu, Jian Yang, Yiming Li, Wenting Chen, Dongrui Liu, Lukas Vierling, Zhiheng Xi, Haobo Fu, Wenxuan Wang, Jitao Sang, Zhengyan Shi, Chi-Min Chan, Eugenie Shi, Simin Li, Juncheng Li, Wei Ji, Dong Li, Jun Song, Yinpeng Dong, Jie Fu, Bo Zheng, Min Yang, Yike Guo, Philip Torr, Zhongyuan Wang, Yaodong Yang✉, Tiejun Huang✉, Ya-Qin Zhang✉, Hongjiang Zhang✉, Andrew Yao✉, 2025. AI Deception: Risks, Dynamics, and Controls. *arXiv Preprint: 2511.22619*
- [22] Zhengliang Shi, Yiqun Chen, Haitao Li, Weiwei Sun, Shiyu Ni, Yougang Lyu, Run-Ze Fan, Bowen Jin, Yixuan Weng, Minjun Zhu, Qiuqie Xie, Xinyu Guo, Qu Yang, Jiayi Wu, Jujia Zhao, Xiaqiang Tang, Xinbei Ma, Cunxiang Wang, Jiaxin Mao, Qingyao Ai, **Jen-tse Huang**, Wenxuan Wang, Yue Zhang, Yiming Yang, Zhaopeng Tu✉, Zhaochun Ren✉, 2025. Deep Research: A Systematic Survey. *arXiv Preprint: 2512.02038*
- [21] Ying Xiao, Zhenpeng Chen✉, **Jen-tse Huang**, Wenting Chen, Yepang Liu, Kezhi Li, Mohammad Reza Mousavi, Richard Dobson, Jie M. Zhang, 2025. Bias in Large AI Models for Medicine and Healthcare: Survey and Challenges. *Preprints.org Preprint: 202511.1838*

- [20] Shuqing Li, Jiayi Yan \*, Chenyu Niu \*, **Jen-tse Huang**, Yun Peng, Wenzuan Wang ✉, Yepang Liu, Michael R Lyu, 2025. ComboBench: Can LLMs Manipulate Physical Devices to Play Virtual Reality Games? *arXiv Preprint: 2510.24706*
- [19] Youliang Yuan, Qiuyang Mang, Jingbang Chen, Hong Wan, Xiaoyuan Liu, Junjielong Xu, **Jen-tse Huang**, Wenzuan Wang, Wenxiang Jiao, Pinjia He ✉, 2025. Curing Miracle Steps in LLM Mathematical Reasoning with Rubric Rewards. *arXiv Preprint: 2510.07774*
- [18] Zhengliang Shi, Ruotian Ma, **Jen-tse Huang**, Xinbei Ma, Xingyu Chen, Mengru Wang, Qu Yang, Yue Wang, Fanghua Ye, Ziyang Chen, Shanyi Wang, Cixing Li, Wenzuan Wang, Zhaopeng Tu ✉, Xiaolong Li, Zhaochun Ren ✉, Linus, 2025. Social Welfare Function Leaderboard: When LLM Agents Allocate Social Welfare. *arXiv Preprint: 2510.01164*
- [17] Xinbei Ma, Ruotian Ma, Xingyu Chen, Zhengliang Shi, Mengru Wang, **Jen-tse Huang**, Qu Yang, Wenzuan Wang, Fanghua Ye, Qingxuan Jiang, Mengfei Zhou, Zhuosheng Zhang ✉, Rui Wang, Hai Zhao, Zhaopeng Tu ✉, Xiaolong Li, Linus, 2025. The Hunger Game Debate: On the Emergence of Over-Competition in Multi-Agent Systems. *arXiv Preprint: 2509.26126*
- [16] Jiaxu Zhou \*, **Jen-tse Huang** \*, Xuhui Zhou, Man Ho Lam, Xintao Wang, Hao Zhu, Wenzuan Wang ✉, Maarten Sap ✉, 2025. The PIMMUR Principles: Ensuring Validity in Collective Behavior of LLM Societies. *arXiv Preprint: 2509.18052*
- [15] Bingkang Shi, **Jen-tse Huang**, Guoyi Li, Xiaodan Zhang ✉, Zhongjiang Yao ✉, 2025. FairGamer: Evaluating Biases in the Application of Large Language Models to Video Games. *arXiv Preprint: 2508.17825*
- [14] Yumeng Wang \*, Zhiyuan Fan \*, Jiayu Liu \*, **Jen-tse Huang**, Yi R. Fung ✉, 2025. Diversity-Enhanced Reasoning for Subjective Questions. *arXiv Preprint: 2507.20187*
- [13] Ada Chen \*, Yongjiang Wu \*, Junyuan Zhang \*, Jingyu Xiao, Shu Yang, **Jen-tse Huang**, Kun Wang, Wenzuan Wang ✉, Shuai Wang, 2025. JARVIS or Ultron? A Survey on the Safety and Security Threats of Computer-Using Agents. *arXiv Preprint: 2505.10924*
- [12] **Jen-tse Huang** ✉, Kaiser Sun, Wenzuan Wang, Mark Dredze, 2025. On the Failure of Latent State Persistence in Large Language Models. *arXiv Preprint: 2505.10571*
- [11] Kun Wang \*, Guibin Zhang \*, Zhenhong Zhou ✉, Jiahao Wu ✉, Miao Yu, Shiqian Zhao, Chenlong Yin, Jinhu Fu, Yibo Yan, Hanjun Luo, Liang Lin, Zhihao Xu, Haolang Lu, Xinye Cao, Xinyun Zhou, Weifei Jin, Fanci Meng, Shicheng Xu, Junyuan Mao, Yu Wang, Hao Wu, Minghe Wang, Fan Zhang, Junfeng Fang, Wenjie Qu, Yue Liu, Chengwei Liu, Yifan Zhang, Qiankun Li, Chongye Guo, Yalan Qin, Zhaoxin Fan, Kai Wang, Yi Ding, Donghai Hong, Jiaming Ji, Yingxin Lai, Zitong Yu, Xinfeng Li, Yifan Jiang, Yanhui Li, Xinyu Deng, Junlin Wu, Dongxia Wang, Yihao Huang, Yufei Guo, **Jen-tse Huang**, Qiufeng Wang, Xiaolong Jin, Wenzuan Wang, Dongrui Liu, Yanwei Yue, Wenke Huang, Guancheng Wan, Heng Chang, Tianlin Li, Yi Yu, Chenghao Li, Jiawei Li, Lei Bai, Jie Zhang, Qing Guo, Jingyi Wang, Tianlong Chen, Joey Tianyi Zhou, Xiaojun Jia, Weisong Sun, Cong Wu, Jing Chen, Xuming Hu, Yiming Li, Xiao Wang, Ningyu Zhang, Luu Anh Tuan, Guowen Xu, Jiaheng Zhang, Tianwei Zhang, Xingjun Ma, Jindong Gu, Liang Pang, Xiang Wang, Bo An, Jun Sun, Mohit Bansal, Shirui Pan, Lingjuan Lyu, Yuval Elovici, Bhavya Kailkhura, Yaodong Yang, Hongwei Li, Wenyuan Xu, Yizhou Sun, Wei Wang, Qing Li, Ke Tang, Yu-Gang Jiang, Felix Juefei-Xu, Hui Xiong, Xiaofeng Wang, Dacheng Tao, Philip S. Yu, Qingsong Wen, Yang Liu, 2025. A Comprehensive Survey in LLM(-Agent) Full Stack Safety: Data, Training and Deployment. *arXiv Preprint: 2504.15585*
- [10] Haoxuan Li, Mingyu Derek Ma, **Jen-tse Huang**, Zhaotian Weng, Wei Wang, Jieyu Zhao, 2025. Bi-asInspector: Detecting Bias in Structured Data through LLM Agents. *arXiv Preprint: 2504.04855*
- [9] Ziyi Liu, Priyanka Dey, Zhenyu Zhao, **Jen-tse Huang**, Rahul Gupta, Yang Liu, Jieyu Zhao, 2025. Can LLMs Grasp Implicit Cultural Values? Benchmarking LLMs' Metacognitive Cultural Intelligence with CQ-Bench. *arXiv Preprint: 2504.01127*
- [8] Xiaoying Zhang ✉, Da Peng, Yipeng Zhang, Zonghao Guo ✉, Chengyue Wu, **Jen-tse Huang**, Chi Chen, Wei Ke, Helen Meng ✉, Maosong Sun, 2025. Will Pre-Training Ever End? A First Step

- [7] **Jen-tse Huang**, Dasen Dai, Jen-Yuan Huang, Youliang Yuan, Xiaoyuan Liu, Wenxuan Wang ☐, Wenxiang Jiao, Pinjia He, Zhaopeng Tu, Haodong Duan ☐, 2025. Human Cognitive Benchmarks Reveal Foundational Visual Gaps in MLLMs. *arXiv Preprint: 2502.16435*
- [6] Yongkang Du, **Jen-tse Huang**, Jieyu Zhao, Lu Lin, 2025. FairCoder: Evaluating Social Bias of LLMs in Code Generation. *arXiv Preprint: 2501.05396*
- [5] 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*
- [4] 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*
- [3] 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*
- [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 Preprint: 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 Preprint: 2301.08745*

#### MENTORSHIP

---

- [14] **Jiaxu Zhou**, CUHK 2027 Undergrad, Next: -.
- [13] **Dasen Dai**, CUHK 2027 Undergrad, Next: -.
- [12] **Tailin Jin**, THU 2027 Undergrad, Next: -.
- [11] **Zixi Chen**, PKU 2027 Undergrad, Next: -.
- [10] **Jiantong Qin**, CUHK 2026 Undergrad, Next: -.
- [9] **Jiaqi Hu**, CUHK 2026 Undergrad, Next: -.
- [8] **Chang Chen**, CUHK 2025 Undergrad, Next: -.
- [7] **Yuhang Yan**, CUHK 2025 Undergrad, Next: Cornell Tech Master.
- [6] **Jingyuan Huang**, CUHK 2025 Undergrad, Next: UGA PhD.
- [5] **Linqi Liu**, CUHK 2025 Undergrad, Next: CMU Master.
- [4] **Donald (Man Ho) Lam**, CUHK 2024 Undergrad, Next: CUHK PhD, honored with CUHK Vice-Chancellor's PhD Scholarship Scheme 2024.
- [3] **Derek (Man Tik) Ng**, CUHK 2024 Undergrad, Next: CUHK RA.
- [2] **Eric John Li**, CUHK 2024 Undergrad, Next: CUHK PhD, honored with Hong Kong PhD Fellowship Scheme (HKPFS) 2025.
- [1] **Billy (Hui Tung) Tse**, CUHK 2024 Undergrad, Next: Co-Founder of SmartQuest.