









# JEN-TSE (JAY) HUANG

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

## EXPERIENCE

---

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

## EDUCATION

---

Ph.D. in Computer Science, The Chinese University of Hong Kong	<i>Aug. 2020 - Dec. 2024</i>
B.Sc. in Computer Science, Peking University	<i>Sep. 2015 - Jul. 2019</i>











## CONFERENCE PAPERS

\* equal contribution     corresponding author

- [32] Man Ho Lam, Chaozheng Wang , **Jen-tse Huang**, Michael R. Lyu, 2025. CodeCrash: Stress Testing LLM Reasoning under Structural and Semantic Perturbations. In *Advances in Neural Information Processing Systems 38*. (NeurIPS'25)
- [31] 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)
- [30] 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*. (EMNLP Main'25)
- [29] **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*. (EMNLP Main'25)
- [28] 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*. (EMNLP Main'25)
- [27] 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*. (EMNLP Main'25)
- [26] **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*. (EMNLP Findings'25)
- [25] 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. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 3149-3167. (ACL Main'25)
- [24] 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)

- [23] 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. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 17825-17846. (ACL Main'25)
- [22] Wenxuan Wang, Kuiyi Gao, 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. In *Findings of the Association for Computational Linguistics ACL 2025*. pp. 10940-10957. (ACL Findings'25)
- [21] **Jen-tse Huang**, Jiaxu Zhou, Tailin Jin, Xuhui Zhou, Zixi Chen, Wenxuan Wang ✉, Youliang Yuan, Michael R. Lyu, Maarten Sap ✉, In *Proceedings of the Forty-Second International Conference on Machine Learning*. (ICML'25)
- [20] Xintao Wang ✉, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, **Jen-tse Huang**, Siyu Yuan, Haoran Guo, Jiangjie Chen, Wei Wang, Yanghua Xiao, Shuchang Zhou, 2025. CoSER: Coordinating LLM-Based Persona Simulation of Established Roles. In *Proceedings of the Forty-Second International Conference on Machine Learning*. (ICML'25)
- [19] Xuhui Zhou, Zhe Su, Sophie Feng, Jiaxu Zhou, **Jen-tse Huang**, Hsien-Te Kao, Spencer Lynch, Svitlana Volkova, Tongshuang Sherry Wu, Anita Woolley, Hao Zhu, Maarten Sap, 2024. SOTOPIA-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, 2024. Competing Large Language Models in Multi-Agent Gaming Environments. In *Proceedings of the Thirteenth International Conference on Learning Representations*. pp. 1-43. (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] **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)
- [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*. pp. 6718-6746. (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*. pp. 2124-2155. (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*. pp. 3781-3789. (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)

---

#### 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. In *Neurocomputing*, vol. 615, no. 128833.

---

#### PREPRINT PAPERS

- [14] 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*.

- [13] Ada Chen \*, Yongjiang Wu \*, Junyuan Zhang \*, Shu Yang, **Jen-tse Huang**, Kun Wang, Wenxuan Wang ✉, Shuai Wang, 2025. A Survey on the Safety and Security Threats of Computer-Using Agents: JARVIS or Ultron? *arXiv Preprint: 2505.10924*.
- [12] **Jen-tse Huang** ✉, Kaiser Sun, Wenxuan Wang, Mark Dredze, 2025. LLMs Do Not Have Human-Like Working Memory. *arXiv Preprint: 2505.10571*.
- [11] Kun Wang \*, Guibin Zhang \*, Zhenhong Zhou ✉, Jiahao Wu ✉, ..., **Jen-tse Huang**, ..., 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 Toward Next-Generation Foundation MLLMs via Self-Improving Systematic Cognition. *arXiv Preprint: 2503.12303*.
- [7] **Jen-tse Huang**, Dasen Dai, Jen-yuan Huang, Youliang Yuan, Xiaoyuan Liu, Wenxuan Wang ✉, Wenxiang Jiao, Pinjia He, Zhaopeng Tu ✉, 2025. VisFactor: Benchmarking Fundamental Visual Cognition in Multimodal Large Language Models. *arXiv Preprint: 2502.16435*.
- [6] Yongkang Du, **Jen-tse Huang**, Jieyu Zhao, Lu Lin, 2025. FairCode: 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 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*.

#### AWARD AND HONOR

- |   |  |
|---|--|
| <ul style="list-style-type: none"> <li>• Outstanding Reviewer</li> <li>• Top Reviewer [1304/15160, 8.6%]</li> </ul> | <i>EMNLP 2024</i><br><i>NeurIPS 2024</i> |
|---|--|

#### SERVICE

- |   |  |
|---|--|
| <ul style="list-style-type: none"> <li>• Conference Reviewer: ICML'25; ICLR'25,26; NeurIPS'24,25; CVPR'24,25,26; ICCV'25; ACL'23,25; EMNLP'23,24,25; NAACL'25' NLPCC'25;</li> <li>• Journal Reviewer: Nature Human Behavior; TMLR;</li> <li>• Organizer: NENLP'25;</li> <li>• Teaching Assistant: Software Engineering (CSCI3100) in CUHK</li> <li>• Teaching Assistant: Discrete Math for Engineers (ENGG2440A) in CUHK</li> </ul> | <br><br><br><br><i>2021, 2022</i><br><i>2020</i> |
|---|--|