Wenxuan **Wang**

jwxwang@gmail.com | https://jarviswang94.github.io/

Summary

I am a **postdoc** researcher at **the Hong Kong University of Science and Technology**, advised by Prof. Shuai Wang. I received my Ph.D. degree at The Chinese University of Hong Kong, advised by Prof. Michael R. Lyu. I was a visiting scholar at UCLA NLP Group, worked with Prof. Nanyun Peng and Harvard Medical School, worked with Prof. Xiang Li.

My research interests are the **trustworthy AI models**, such as Large Language Models, Multi-Modal AI Models, Code Intelligent Models, and Medical Intelligent Models. My works have over 2000 citations and an h-index of 21. My GitHub projects have received over 3000 stars.

Education

The Chinese University of Hong Kong — *Hong Kong*

PhD's Student in Computer Science and Engineering — *2020 - 2024*

Huazhong University of Science and Technology — *Wuhan, Hubei, China*

B.E. in Computer Science and Technology — *2013 - 2017*

Experience

Hong Kong University of Science and Technology — *Hong Kong, China*

PostDoc Researcher — *2024.9 - Now*

- Security and Safety of LLMs

UCLA NLP Group — *Los Angeles, CA, USA*

Visiting Researcher — *2023 - 2024*

- Evaluation of Large Language Models and Multi-modal AI Models

Tencent AI Lab — *Shenzhen, Guangdong, China*

Research Intern — *2019 - 2023*

- Safety and Reliability of Language Generation Models

Turing Robot Inc. — *Beijing, China*

Research and Development Engineer — *2017 - 2019*

- Multi-modal AI

Selected Publication

Apathetic or Empathetic? Evaluating LLMs' Emotional Alignment with Humans — *NeurIPS 2024*

JEN-TSE HUANG, MAN HO LAM, ERIC JOHN LI, SHUJIE REN, **WENXUAN WANG (CORRESPONDING)**, WENXIANG JIAO, ZHAOPENG TU, MICHAEL R. LYU — *(CCF A)*

the Thirty-Eighth Annual Conference on Neural Information Processing Systems

On the Reliability of Psychological Scales on Large Language Models — *EMNLP 2024*

JEN-TSE HUANG, WENXIANG JIAO, MAN HO LAM, ERIC JOHN LI, **WENXUAN WANG (CORRESPONDING)**, MICHAEL R. LYU — *(CAAI A, CCF B)*

The 2024 Conference on Empirical Methods in Natural Language Processing

LogicAsker: Triggering the Logical Reasoning Failures in Large Language Models — *EMNLP 2024*

YUXUAN WAN*, **WENXUAN WANG (CO-FIRST)***, YILIU YANG, PINJIA HE, WENXIANG JIAO, MICHAEL R. LYU — *(CAAI A, CCF B)*

The 2024 Conference on Empirical Methods in Natural Language Processing

New Job, New Gender? Measuring the Social Bias in Image Generation Models — *ACM MM 2024 (Oral)*

WENXUAN WANG, HAONAN BAI, JEN-TSE HUUANG, JINGYUAN HUANG , HAOYI QIU, NANYUN PENG, MICHAEL R. LYU — *(CCF A)*

ACM Multimedia 2024 (Oral)

Not All Countries Celebrate Thanksgiving: On the Cultural Dominance in Large Language Models — *ACL 2024*

WENXUAN WANG, WENXIANG JIAO, RUYI DAI, JINGYUAN HUANG, JEN-TSE HUANG, ZHAOPENG TU, MICHAEL R. LYU — *(CCF A)*

The Annual Meeting of the Association for Computational Linguistics

All Languages Matter! A Multilingual Safety Benchmark for Large Language Models *ACL 2024 Findings*

WENXUAN WANG, ZHAOPENG TU, CHANG CHEN, YOULIANG YUAN, JEN-TSE HUANG, WENXIANG JIAO, MICHAEL R. LYU *(CCF A)*

Findings of The Annual Meeting of the Association for Computational Linguistics

Understanding and Mitigating the Uncertainty in Zero-Shot Translation *TASLP*

WENXUAN WANG, WENXIANG JIAO, SHUO WANG, ZHAOPENG TU, MICHAEL R. LYU *(JCR Q1, SCI Q1, CAAI A)*

The IEEE/ACM Transactions on Audio, Speech, and Language Processing

On the Shortcut Learning in Multilingual Neural Machine Translation *Neurocomputing*

WENXUAN WANG, WENXIANG JIAO, JEN-TSE HUANG, ZHAOPENG TU, MICHAEL R. LYU *(JCR Q2, SCI Q1)*

Neurocomputing

A Picture is Worth a Thousand Toxic Words: A Metamorphic Testing Framework for Content Moderation Software *ASE 2023*

WENXUAN WANG, JINGYUAN HUANG, CHANG CHEN, PINJIA HE, JIAZHEN GU, MICHAEL R. LYU *(CCF A)*

The IEEE/ACM International Conference on Automated Software Engineering

BiasAsker: Measuring the Bias in Conversational AI System *FSE 2023*

YUXUAN WAN*, **WENXUAN WANG* (CO-FIRST)**, PINJIA HE, JIAZHEN GU, HAONAN BAI, MICHAEL R. LYU *(CCF A)*

The ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering

Validating Multimedia Content Moderation Software via Semantic Fusion *ISSTA 2023*

WENXUAN WANG, JINGYUAN HUANG, CHANG CHEN, JIAZHEN GU,JIANPING ZHANG, WEIBIN WU, PINJIA HE, MICHAEL R. LYU *(CCF A)*

The ACM SIGSOFT International Symposium on Software Testing and Analysis

MTTM: Metamorphic Testing for Textual Content Moderation Software *ICSE 2023*

WENXUAN WANG, JEN-TSE HUANG, WEIBIN WU, JIANPING ZHANG, YIZHAN HUANG, SHUQING LI, PINJIA HE, MICHAEL R. LYU *(CCF A)*

The IEEE/ACM International Conference on Software Engineering

Understanding and Improving Sequence-to-Sequence Pretraining for Neural Machine Translation *ACL 2022*

WENXUAN WANG, WENXIANG JIAO, YONGCHANG HAO, XING WANG, SHUMING SHI, ZHAOPENG TU, MICHAEL R. LYU *(CCF A)*

The Annual Meeting of the Association for Computational Linguistics

Rethinking the Value of Transformer Components *COLING 2020*

WENXUAN WANG, ZHAOPENG TU *(CCF B)*

The International Conference on Computational Linguistics

FPETS: Fully Parallel End-to-end Text-to-speech System *AAAI 2020 Oral*

DABIAO MA*, ZHIBA SU*, **WENXUAN WANG* (CO-FIRST)**, YUHAO LU *(CCF A)*

AAAI Conference on Artificial Intelligence

Who is ChatGPT? Benchmarking LLMs' Psychological Portrayal Using PsychoBench *ICLR 2024 Oral*

JEN-TSE HUANG, **WENXUAN WANG**, ERIC JOHN LI, MAN HO LAM, SHUJIE REN, YOULIANG YUAN, WENXIANG JIAO, ZHAOPENG TU, MICHAEL R. LYU *(CAAI A)*

The Twelfth International Conference on Learning Representations

Gpt-4 is Too Smart to be Safe: Stealthy Chat with LLMs via Cipher *ICLR 2024*

YOULIANG YUAN, WENXIANG JIAO, **WENXUAN WANG**, JEN-TSE HUANG, PINJIA HE, SHUMING SHI, ZHAOPENG TU *(CAAI A)*

The Twelfth International Conference on Learning Representations

Boosting Adversarial Transferability by Block Shuffle and Rotation *CVPR 2024*

KUNYU WANG, XUANRAN HE, **WENXUAN WANG**, XIAOSEN WANG *(CCF A)*

The IEEE/CVF Conference on Computer Vision and Pattern Recognition

Does ChatGPT Know that It Does Not Know? Evaluating the Black-Box Calibration of ChatGPT. *COLING 2024*

YOULIANG YUAN, **WENXUAN WANG**, QINGSHUO GUO, YIMING XIONG, CHIHAO SHEN, PINJIA HE *(CCF B)*

International Conference on Computational Linguistics

Generative Type Inference for Python *ASE 2023*

YUN PENG, CHAOZHENG WANG, **WENXUAN WANG**, CUIYUN GAO, MICHAEL R. LYU *(CCF A)*

The IEEE/ACM International Conference on Automated Software Engineering (**ACM SIGSOFT Distinguished Paper Award**)

Constructing Effective In-Context Demonstration for Code Intelligence Tasks: An Empirical Study *ASE 2023*

SHUZHENG GAO, XIN-CHENG WEN, CUIYUN GAO, **WENXUAN WANG**, MICHAEL R. LYU *(CCF A)*

The IEEE/ACM International Conference on Automated Software Engineering

Improving the Transferability of Adversarial Samples by Path-Augmented Method *CVPR 2023*

JIANPING ZHANG, JEN-TSE HUANG, **WENXUAN WANG**, YICHEN LI, WEIBIN WU, XIAOSEN WANG, YUXIN SU, MICHAEL R. LYU *(CCF A)*

The IEEE/CVF Conference on Computer Vision and Pattern Recognition

Parrot: Translating during chat using large language models *Findings of EMNLP 2023*

WENXIANG JIAO, JEN-TSE HUANG, **WENXUAN WANG**, XING WANG, ZHAOPENG TU *(CAAI A, CCF B)*

Findings of The Conference on Empirical Methods in Natural Language Processing

LFAA: Crafting Transferable Targeted Adversarial Examples with Low-Frequency Perturbations *ECAI 2023*

KUNYU WANG, JULUAN SHI, **WENXUAN WANG** *(CCF B)*

26th European Conference on Artificial Intelligence

Improving Adversarial Transferability via Neuron Attribution-Based Attacks *CVPR 2022*

JIANPING ZHANG, WEIBIN WU, JEN-TSE HUANG, YIZHAN HUANG, **WENXUAN WANG**, YUXIN SU, MICHAEL R. LYU *(CCF A)*

The IEEE/CVF Conference on Computer Vision and Pattern Recognition

AEON: A Method for Automatic Evaluation of NLP Test Cases *ISSTA 2022*

JEN-TSE HUANG, JIANPING ZHANG, **WENXUAN WANG**, PINJIA HE, YUXIN SU, MICHAEL R. LYU *(CCF A)*

The ACM SIGSOFT International Symposium on Software Testing and Analysis

Revisiting, Benchmarking and Exploring API Recommendation: How Far Are We? *TSE 2021*

YUN PENG, SHUQING LI, WENWEI GU, YICHEN LI, **WENXUAN WANG**, CUIYUN GAO, MICHAEL R. LYU *(CCF A)*

IEEE Transactions on Software Engineering