HONGRU WANG

Emails: hongru.carrywang@gmail.com / hrwise98@gmail.com
Homepage | Ph.D. Thesis | Thesis Slides | Google Scholar (2100+ citations)

RESEARCH STATEMENT

My research focus revolves around ***Theory of Agent*** (***ToA***), which unifying internal reasoning and external acting (a.k.a., two major behaviors) of agent as two epistemically alternative ways to model the internal world stored in the parametric space and external digital world. My long-term objective is to achieve the *impossible triangle* between safety, personalization and autonomy of language agent to learn from interactions internally or externally.

EDUCATION

University of Edinburgh **2025/09 - Now**
Postdoc (Research Associate) supervised by Prof. Amos Storkey and Prof. Jeff Z. Pan
Self-evolving Agents / Long-horizon Reasoning / Tool Learning

The Chinese University of Hong Kong **2021/09 - 2025/09**
Ph.D. supervised by Prof. Kam-Fai Wong (ACL Fellow) GPA: 3.82/4.00
Dialogue System, Large Language Models, Tool Learning

The Chinese University of Hong Kong **2019/09 - 2020/11**
M.Eng. Computer Science and Engineering GPA: 3.60/4.00
Data Mining, Big Data Technology, Machine Learning

Communication University of China **2015/09 - 2019/09**
B.Eng. Computer Science and Technology GPA: 3.27/4.00
Computer Operation System, Computer Networks, Data Structures and Algorithms

EXPERIMENCE

Visiting Student University of Illinois Urbana-Champaign **2024/12 - 2025/6**
BlenderLab (Mentor: Prof. Heng Ji (ACL Fellow))

Research Intern Seed-LLM-Harizon (Doubao) **2024/08 - 2024/12**
ByteDance (SZ) (Mentor: Dr. Deng Cai, Dr. Wanjun Zhong)

Visiting Student University of Edinburgh **2024/04 - 2024/11**
EdinburghNLP (Mentor: Prof. Jeff Z. Pan)

Research Assistant MoE Key Lab of High Confidence Software Technologies **2019/12 - 2021/07**
CUHK (Mentor: Prof. Kam-Fai Wong)

(CO-)FIRST/CORRESPONDING AUTHOR CONFERENCES

22. From Word to World: Can LLM be Implicit Text-based World Model?
Yixia Li, Hongru WANG†, Jiahao Qiu, Zhenfei Yin et al., Guanhua Chen, Heng Ji
ACL 2026 LLM World Model † *denotes corresponding author*

21. Rethinking the Role of Entropy in Optimizing Tool-Use Behaviors for Large Language Model Agents
Zeping Li, Hongru WANG†, Yiwen Zhao, Guanhua Chen, Yixia Li et al., Zhenfei Yin
ACL 2026 LLM Tool Learning

20. Self-Sum: Teaching Agent Itself to Decide When and What to Summarize
Hongru WANG, Rui Wang, Jushi Kai et al., Xiaoteng Ma, Jeff Z. Pan, Amos Storkey
Findings of **ACL 2026** LLM Tool Learning

19. DecisionFlow: Advancing Large Language Model as Principled Decision Maker
Xiusi Chen*, Shanyong Wang*, Cheng Qian*, Hongru WANG*, Peixuan Han, and Heng Ji
Findings of **EMNLP 2025** LLM Tool Learning * *denotes equal contribution*

18. SafeToolBench: Pioneering a Prospective Benchmark to Evaluating Tool Utilization Safety in LLMs
Hongfei Xia*, Hongru WANG*, Zeming Liu, Qian Yu, Yuhang Guo, and Haifeng Wang
Findings of **EMNLP 2025** LLM Tool Learning

17. UAlign: Leveraging Uncertainty Estimations for Factuality Alignment on Large Language Models
Boyang XUE, Fei Mi, Qi Zhu, Hongru WANG†, Rui Wang, et al., Kam-Fai Wong
ACL 2025 LLM

16. Self-Reasoning Language Models: Unfold Hidden Reasoning Chains with Few Reasoning Catalyst
Hongru WANG, Deng Cai, Wanjun Zhong, Shijue Huang, Jeff Z. Pan, Zeming Liu, Kam-Fai Wong
Reasoning and Planning for Large Language Models of **ICLR 2025**
Findings of **ACL 2025** LLM (data synthesise to self-improve itself for LLMs)

15. Rethinking Stateful Tool Use in Multi-Turn Dialogues: Benchmarks and Challenges
Hongru WANG, Wenyu Huang, et al., Zeming Liu, Jeff Z. Pan, Kam-Fai Wong
Findings of **ACL 2025** LLM Tool Learning

14. MlingConf: A Comprehensive Study of Multilingual Confidence Estimation on LLMs
Boyang XUE*, Hongru WANG*, Rui Wang, et al., Wenxuan Zhang, Kam-Fai Wong
Findings of **ACL 2025** LLM

13. ToolSpectrum: Towards Personalized Tool Utilization for Large Language Models
Zihao Cheng*, Hongru WANG*, Zeming Liu, Yuhang Guo, et al., Yunhong Wang, Haifeng Wang
Findings of **ACL 2025** LLM Tool Learning

12. TransBench: Breaking Barriers for Transferable GUI Agents in Dynamic Digital Environments
Yuheng Lu*, Qian Yu*, Hongru WANG*, Zeming Liu, Wei Su, et al., Yunhong Wang, Haifeng Wang
Findings of **ACL 2025** LLM Agent

11. Self-DC: When to Reason and When to Act? Self Divide-and-Conquer for Compositional ..
Hongru WANG, Boyang Xue, Baohang Zhou, et. al, Kam-Fai Wong
NAACL 2025 LLM , Tool Learning **Oral** Blog

10. AppBench: Planning of Multiple APIs from Various APPs for Complex User Instruction
Hongru WANG, Rui Wang, Boyang XUE, et. al, Jeff Z. Pan, Kam-Fai Wong
EMNLP 2024 LLM , Tool Learning (Apple Intelligence)

9. TPE: Towards Compositional Reasoning over Conceptual Tools with Multi-persona Collaboration
Hongru WANG, Huimin Wang, Lingzhi Wang, et. al, Kam-Fai Wong
NLPCC 2024 LLM , Tool Learning (Meta-reasoning theory / Cognitive tools)

8. Enhancing Large Language Models Against Inductive Instructions with Dual-critique Prompting
Rui Wang*, Hongru WANG*, Fei Mi, Boyang XUE, Yi Chen, Kam-Fai Wong, Ruifeng Xu
NAACL 2024 LLM , Safety

7. UniRetriever: Multi-task Candidates Selection for Various Context-Adaptive Conversational Retrieval
Hongru WANG, Boyang Xue, Baohang Zhou, et. al, Kam-Fai Wong
LREC-COLING 2024 Dialogue System , Tool Learning

6. M3Sum: A Novel Unsupervised Language-guided Video Summarization
Hongru WANG, Baohang Zhou, Zhengkun Zhang, Yiming Du, David Ho, Kam-Fai Wong
ICASSP 2024 LLM , MM

5. Large Language Models as Source Planner for Personalized Knowledge-grounded Dialogue
Hongru WANG, Minda Hu, Yang Deng, et. al, Irwin King, Kam-Fai Wong
Findings of **EMNLP 2023** Dialogue System , LLM **Best Paper Award 😄 @Doctoral Forum**

4. Cue-CoT: Chain-of-thought Prompting for Responding to In-depth Dialogue Questions with LLMs
Hongru WANG*, Rui Wang*, Fei Mi, et. al, Ruifeng Xu and Kam-Fai Wong
Findings of **EMNLP 2023** Dialogue System , LLM Blog

3. MCML: A Novel Memory-based Contrastive Meta-Learning Method for Few Shot Slot Tagging
Hongru WANG, Zezhong Wang, Wai-Chung Kwan, Kam-Fai Wong
IJCNLP-AACL 2023 Dialogue System

2. Integrating Pretrained Language Model for Dialogue Policy Learning
Hongru WANG, Huimin Wang, Zezhong Wang and Kam-Fai Wong
ICASSP 2022 Dialogue System , RLHF (probably first work of RLHF)

1. CUHK at SemEval-2020 Task 4: CommonSense Explanation, Reasoning and Prediction ..
Hongru WANG, Xiangru Tang, Sunny Lai, et. al, Gabriel Pui Cheong Fung and Kam-Fai Wong
SemEval of **COLING 2020**. Commonsense

(CO-)FIRST/CORRESPONDING AUTHOR JOURNAL

4. AdaCtrl: Towards Adaptive and Controllable Reasoning via Difficulty-Aware Budgeting
Shijue Huang*, Hongru WANG*, Wanjun Zhong, et al, Bowen Cao, Yi R. (May) Fung
Transactions on Machine Learning Research 2026
Teaching model to reason efficiently

3. A Survey of Self-Evolving Agents: What, When, How, and Where to Evolve on the Path to Artificial Super Intelligence
Huan-ang Gao et al., Hongru WANG*†, et al., Qingyun Wu, Heng Ji, Mengdi Wang†
Transactions on Machine Learning Research 2026 Self-Evolving Agents
First comprehensive survey on self-evolving agents; served as one of two corresponding authors.

2. KddRES: A Multi-level Knowledge-driven Dialogue .. Towards Customized Dialogue System
Hongru WANG, Wai-Chung Kwan, Min Li, Zimo Zhou and Kam-Fai Wong
Computer Speech and Language 2024 (Tsinghua B, JCR Q2, IF: 4.3) Dialogue System

1. A Survey on Recent Advances and Challenges in Reinforcement Learning Methods ..
Wai-Chung Kwan*, Hongru WANG*, Huimin Wang and Kam-Fai Wong
Machine Intelligence Research 2023 (JCR Q1, IF: 6.4) Dialogue System , RLAIF

NON-FIRST AUTHOR CONFERENCES/JOURNALS

26. Perception-aware Policy Optimization for Multimodal Reasoning
Zhenhailong Wang et al., Hongru WANG, et al., Fei Huang, Heng Ji
ICLR 2026 Multimodal Reasoning

25. RM-R1: Reward Modeling as Reasoning
Xiusi Chen, et al., Hongru WANG, et al., Tong Zhang, Hanghang Tong, Heng Ji
ICLR 2026 **Reward modeling as reasoning** Blog

24. ToolRL: Reward is All Tool Learning Needs
Cheng Qian, et al, Hongru WANG, Xiusi Chen, Dilek Hakkani-Tür, Gokhan Tur, Heng Ji
NeurIPS 2025 **A comprehensive analysis of reward design for tool learning** Blog

23. The Landscape of Agentic Reinforcement Learning for LLMs: A Survey
Guibin Zhang et al., Hongru WANG, et al., Jun Wang, Shuicheng YAN, Philip Torr, Lei Bai
Transactions on Machine Learning Research 2026 Agentic Reinforcement Learning

22. Can LLMs Evaluate Complex Attribution in QA? Automatic Benchmarking using Knowledge Graphs
Nan Hu, Jiaoyan Chen, Yike Wu, Guilin Qi, Hongru WANG, et al., Jeff Z. Pan
ACL 2025 LLM

21. ImPart: Importance-Aware Delta-Sparsification for Improved Model Compression and Merging ...
Yan Yang, Yixia Li, Hongru WANG, Xuetao Wei, James Jianqiao Yu, Yun Chen, Guanhua Chen
ACL 2025 LLM Model Compression

20. SMART: Self-Aware Agent for Tool Overuse Mitigation
Cheng Qian, Emre Can Acikgoz, Hongru WANG#, et al., Dilek Hakkani-Tür, Gokhan Tur, Heng Ji
Findings of **ACL 2025** LLM Tool Learning # *denotes mentorship*

19. Steering Knowledge Selection Behaviours in LLMs via SAE-Based Representation Engineering
Yu Zhao, et al., Hongru WANG, Xuanli He, Kam-Fai Wong, Pasquale Minervini
NAACL 2025 LLM (tool / knowledge conflicts, detections and control) **Oral**

18. SeqAR: Jailbreak LLMs with Sequential Auto-Generated Characters
Yan Yang, Zeguan Xiao, Xin Lu, Hongru WANG, et al., Guanhua Chen, Yun Chen
NAACL 2025 LLM

17. AutoPSV: Automated Process-Supervised Verifier
Jianqiao Lu, Zhiyang Dou, Hongru WANG, et. al, Zhijiang Guo
NeurIPS 2024 LLM (Meta-reasoning theory)

16. Knowledge Conflicts for LLMs: A Survey
Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru WANG, Yue Zhang, Wei Xu
EMNLP 2024 LLM Blog

15. VLEU: a Method for Automatic Evaluation for Generalizability of Text-to-Image Models
Jingtao Cao, Zhang Zheng, Hongru WANG, Kam-Fai Wong
EMNLP 2024 MM

14. Less is More: Making Smaller Language Models .. Subgraph Retrievers for Multi-hop KGQA
Wenyu Huang, Guancheng Zhou, Hongru WANG, et. al, Mirella Lapata, Jeff Z. Pan
Findings of **EMNLP 2024** LLM (Differentiable Search Index (DSI))

13. Enhancing Biomedical Knowledge RAG with Self-Rewarding Tree Search and PPO
Minda Hu, Licheng Zong, Hongru WANG, et. al, Kam-Fai Wong, Yu Li, Irwin King
Findings of **EMNLP 2024** LLM

12. DPDLLM: A Black-box Framework for Detecting Pre-training Data from Large Language Models
Baohang Zhou, Zezhong WANG, Lingzhi Wang, Hongru WANG, et. al, Kam-Fai Wong
Findings of **ACL 2024** LLM

11. Medical Dialogue: A Survey of Categories, Methods, Evaluation and Challenges
Xiaoming Shi, Zeming Liu, Li Du, Yuxuan Wang, Hongru WANG, et. al, Shaoting Zhang
Findings of **ACL 2024** LLM

10. REGA: Role Prompting Guided Multi-Domain Adaptation for Large Language Models
Rui Wang, Fei Mi, Yi Chen, Boyang XUE, Hongru WANG, Qi Zhu, Kam-Fai Wong, Ruifeng Xu
Findings of **NAACL 2024** LLM , Role Playing

9. SELF-GUARD: Empower the LLM to Safeguard Itself
Zezhong Wang, Fangkai Yang, Lu Wang, Pu Zhao, Hongru WANG, et. al, Kam-Fai Wong
NAACL 2024 LLM , Safety

8. JoTR: A Joint Transformer and Reinforcement Learning Framework for Dialog Policy Learning
Wai-Chung Kwan, Huimin Wang, Hongru WANG, et. al, Kam-Fai Wong
LREC-COLING 2024 Dialogue System

7. MCIL: Multimodal Counterfactual Instance Learning for .. Multimodal Information Extraction
Baohang Zhou, Ying Zhang, Kehui Song, Hongru WANG, Yu Zhao, Xuhui Sui, Xiaojie Yuan
LREC-COLING 2024 LLM , MM

6. ReadPrompt: A Readable Prompting Method for Reliable Knowledge Probing
Zezhong Wang*, Luyao YE*, Hongru WANG, Wai-Chung Kwan, David Ho, Kam-Fai Wong
Findings of **EMNLP 2023** LLM

5. Improving Factual Consistency for Knowledge-Grounded Dialogue Systems via Knowledge ..
Boyang XUE*, Weichao Wang*, Hongru WANG, et. al, Xin Jiang, Qun Liu, Kam-Fai Wong
Findings of **EMNLP 2023** Dialogue System

4. Prompting and Evaluating Large Language Models for Proactive Dialogues ..
Yang Deng, Lizi Liao, Liang CHEN, Hongru WANG, Wenqiang Lei, Tat-Seng Chua
Findings of **EMNLP 2023** LLM

3. Towards Robust Personalized Dialogue Generation via Order-Insensitive Representation ..
Liang Chen, Hongru WANG, Yang Deng, Wai Chung Kwan, Zezhong Wang, Kam-Fai Wong
Findings of **ACL 2023** (short) Dialogue System

2. Retrieval-free Knowledge Injection through Multi-Document Traversal for Dialogue Models
Rui Wang, Jianzhu Bao, Fei Mi, Yi Chen, Hongru WANG, et. al, Kam-Fai Wong, Ruifeng Xu
ACL 2023 Dialogue System

1. DIGAT: Modeling News Recommendation with Dual-Graph Interaction
Zhiming Mao, Jian Li, Hongru WANG, Xingshan Zeng, Kam-Fai Wong
Findings of **EMNLP 2022** Others

WORKSHOP, TUTORIAL AND OTHERS

5. Lifelong Agents: Learning, Aligning, Evolving
Cheng Qian, Emre Can Acikgoz, Hongru WANG, Zhenfei Yin, Manling Li, Yun-Nung Chen, Mengdi Wang, Caiming Xiong
First Workshop on Lifelong Agents at ICLR 2026 Lifelong Agents
Workshop Organizer and Program Chair

4. Analysing the Residual Stream of Language Models Under Knowledge Conflicts
Yu Zhao, Xiaotang Du, et. al, Hongru WANG, Xuanli He, Kam-Fai Wong, Pasquale Minervini
MINT Workshop of **NeurIPS 2024**

3. PerLTQA: A Personal Long-Term Memory Dataset for Memory Classification, Retrieval, and ..
Yiming Du, Hongru WANG, Zhengyi Zhao, et. al, Kam-Fai Wong
SIGHAN Workshop of **ACL 2024**. **Best Paper Award 😄.**

2. OSPC: Detecting Harmful Memes with Large Language Model as a Catalyst
   Jingtao Cao, Zheng Zhang, Hongru WANG, Bin Liang, Hao Wang, Kam-fai Wong
   Online Safety Prize Challenge, **WWW 2024**. **Champion Solution 😄.**

1. Empowering Large Language Models: Tool Learning for Real-World Interaction
   Hongru WANG, Yujia Qin, Yankai Lin, Jeff Z. Pan, Kam-fai Wong
   Tutorial, **SIGIR 2024** LLM , Tool Learning

GRANTS / FUNDING

3. Overseas Research Attachment Programme (50,000 HKD, ORAP 2023-2024)
2. A Knowledge Graph Based Dynamic Video Extractive Summarization System (PRP/054/21FX)
1. "SEVES: Semantic-driven Effective Video Extractive Summarization System" – Technology and Business Development Fund (200,000 HKD, TBF22ENG004)

TEACHING ASSISTANT

SEEM 3450 Engineering Innovation and Entrepreneurship
SEEM 3490 Information Systems Management
SEEM 5730 / ECLT 5910 Information Technology Management

COMPETITIONS & AWARDS

Top 1 at Online Safety Prize Challange, WWW 2024	international
Reaching Out Awards (2022-2023)	school
Top 10 at ICLR 2021 Workshop MLPCP Track 1 (rank 7th)	international
Third Price at SMP2020-ECDT Few-shot Spoken Language Understanding	international
Top 10 at SemEval2020-Task4: CommonSense Detection and Explanations	international
Distinguished Academic Performance Scholarship (2019-2020), CUHK CSE	school
Meritorious Winner, Mathematical Contest In Modeling (2018)	international
A software copyright of "WeCampus WeChat mini-program" (2018SR562540)	national
China National Radio Scholarship	school

COMMUNITY CONTRIBUTION

Program Chair: Lifelong Agents @ ICLR 2026
Area Chair: NeurIPS
Reviewer: ICLR, ICML, ACL/EMNLP/NAACL, AAAI, IJCAI, ARR, TNNLS, ...
Organizer: ToolsMeetLLM@SIGIR2024
Co-founder and Committee: NLP Academic Exchange Platform (NICE) (130,000+ followers, 7,000,000+ views)

REFERENCES

Prof. Kam-Fai Wong, ACL Fellow, Professor, The Chinese University of Hong Kong
Prof. Heng Ji, ACL Fellow, Professor, University of Illinois Urbana-Champaign
Prof. Jeff Z. Pan, Turing Fellow, Professor, University of Edinburgh
Prof. Amos Storkey, Professor, University of Edinburgh
Prof. Irwin King, ACM/IEEE/AAAI Fellow, Professor, The Chinese University of Hong Kong