Jiaying Wu

jiayingw@mit.edu | jiayingwu19.github.io | Google Scholar

Research Overview

My research develops **socially intelligent AI agents** that can (1) **Understand** latent human and social states under uncertainty, (2) **Act** through grounded interventions that support human decisions and complex knowledge workflows, and (3) **Evolve** by adapting as interaction shapes future states and behavior. My long-term goal is to build **long-horizon social intelligence**, enabling AI agents to model evolving humans, anticipate the effects of their actions, and learn safely over extended interaction.

Academic Positions

Massachusetts Institute of Technology — Cambridge, MA, USA
Postdoctoral Researcher — 11/2026 – Present

- Advisor: Paul Liang.
- Research focus: socially intelligent AI agents.

National University of Singapore — Singapore
Research Fellow — 07/2024 – 11/2026

- Advisor: Min-Yen Kan.
- Research focus: socially aware AI for trustworthy information ecosystems.

Education

National University of Singapore — Singapore
Ph.D. in Computer Science — 2020 – 2024

- Advisor: Bryan Hooi.
- Thesis: *Towards Real-World Generalizability of Misinformation Detection*.

Beijing Foreign Studies University — Beijing, China
B.Eng. in Computer Science; Double Degree (B.A.) in English — 2016 – 2020

- Selected for the BFSU Xinmiao Program for Outstanding Undergraduate Students.
- Thesis: *Story Rewriting Based on Counterfactual Reasoning*.

Grants Contributed

Led proposal development and coordinated research execution in collaboration with the listed PIs.

Automating Community Note Generation for Social Media Misinformation Governance 2025 – 2028

- PI: Min-Yen Kan. Amount: SGD 100,000. Singapore MOE Academic Research Fund (AcRF) Tier 1.

Towards Uncertainty-Aware Evidence-Based Misinformation Detection with LLMs 2025 – 2028

- PI: Bryan Hooi. Amount: SGD 100,000. Singapore MOE Academic Research Fund (AcRF) Tier 1.

CheckUp: Grounding Social Media on Healthcare with Science 2024 – 2026

- PI: Min-Yen Kan. Amount: SGD 100,000. Crazy Ideas Seed Grant, NUS School of Computing.

PUBLICATIONS

(* Equal contribution; Research mentees; † Corresponding author. Full list on Google Scholar.)

Selected Preprints

[p.1] Chichen Lin, Huaijin Xie, Wenhan Xie, Weijian Fan, Xiao Han, Yijie Jin, Xiaojie Hao, **Jiaying Wu**†.When Platforms Pay Users to Correct Misinformation: Revenue Incentives and Governance Capacity in Weibo Annotations. *arXiv:2610.18668*.

[p.2] Yibo Li, Jinhang Qiu, Zhi Zheng, Qianyun Guo, **Jiaying Wu**, Shuo Ji, Bryan Hooi. Learn2Play Bench: How Well Do LLM Agents Learn from Experience in Unfamiliar Environments?. *arXiv:2610.08215*.

[p.3] Herun Wan, **Jiaying Wu**†, Minnan Luo†, Zihan Ma, Fanxiao Li, Nancy F. Chen, Min-Yen Kan. Right Answers, Wrong States: Hidden Information Failures in Multi-Agent Collaboration. *arXiv:2610.01244*.

[p.4] Jundong Xu*, Qingchuan Li*, **Jiaying Wu**, Yihuai Lan, Shuyue Stella Li, Huichi Zhou, Bowen Jiang, Lei Wang, Jun Wang, Anh Tuan Luu, Caiming Xiong, Hae Won Park, Bryan Hooi, Zhiyuan Hu. EvoArena: Tracking Memory Evolution for Robust LLM Agents in Dynamic Environments. *arXiv:2606.13681*.

[p.5] Herun Wan, **Jiaying Wu**†, Minnan Luo†, Fanxiao Li, Ningnan Wang, Nancy F. Chen, Min-Yen Kan. The Deliberative Illusion: Diagnosing Factual Attrition and Stance Homogenization in Multi-Agent LLM Deliberation. *arXiv:2606.03032*.

[p.6] Zihang Fu, Haonan Wang, Jian Kang, Kenji Kawaguchi, **Jiaying Wu**†. Lost in Adaptation: Layer-Selective Recovery of Temporal Reasoning in Video-Language Models. *arXiv:2604.11399*.

Conference and Journal Publications

[c.1] Fanxiao Li, **Jiaying Wu**†, Tingchao Fu, Natasha Jaques, Wei Zhou†, Min-Yen Kan. FlowSteer: Prompt-Only Workflow Steering Exposes Planning-Time Vulnerabilities in Multi-Agent LLM Systems. *NeurIPS 2026*.

[c.2] Herun Wan, **Jiaying Wu**†, Minnan Luo†, Fanxiao Li, Zhi Zeng, Min-Yen Kan. The Facade of Truth: Uncovering and Mitigating LLM Susceptibility to Deceptive Evidence. *EMNLP 2026*.

[c.3] Jiecheng Jiang, Herun Wan, Zihang Fu, Hai-Tao Yu, Min-Yen Kan, **Jiaying Wu**†. Why Is This Joke Offensive? Interpreting Offensiveness in Chinese Dark Humor Through Knowledge and Rhetoric. *EMNLP 2026*.

[c.4] Zihang Fu, Fanxiao Li, Jianyang Gu, Haonan Wang, Preslav Nakov, Bryan Hooi, Min-Yen Kan, **Jiaying Wu**†. Better with Experience: Self-Evolving LLM Agents for Evidence-Grounded Health Community Notes. *EMNLP 2026 Findings*.

[c.5] Weida Liang*, Zihang Fu*, Haonan Wang*, Nie Zheng, Yifan Zhang, Yao Tong, Tongyao Zhu, Hao Jiang, Chuang Li, **Jiaying Wu**, Kenji Kawaguchi. From Harm to Help: Turning Reasoning In-Context Demos into Assets for Reasoning LMs. *EMNLP 2026 Findings*.

[c.6] Zhi Zeng, **Jiaying Wu**, Xulang Zhang, Xiangzheng Kong, Herun Wan, Guodong Li, Qika Lin, Minnan Luo, Erik Cambria. Learning from Incomplete Signals: Multi-View Knowledge Distillation for Robust Harmful Micro-Video detection. *Information Fusion, 2026*.

[c.7] Beizhe Hu, Qiang Sheng, Hao Mi, **Jiaying Wu**, Zhengjia Wang, Yuanlong Yu, Danding Wang, Xuming Hu, Juan Cao. EvoFEND: Dual Memory-Driven Self-Evolving Fake News Detection. *KDD 2026*.

[c.8] **Jiaying Wu***, Zihang Fu*, Haonan Wang, Fanxiao Li, Jiafeng Guo, Preslav Nakov, Min-Yen Kan. Beyond the Crowd: LLM-Augmented Community Notes for Governing Health Misinformation. *ACL 2026 (Oral)*.

[c.9] Zhi Zeng, **Jiaying Wu**, Minnan Luo, Di Zhang, Yifei Yang, Xiangzheng Kong, Herun Wan, Zihan Ma. From Detection to Understanding: Multi-Turn Reasoning for Video Misinformation Analysis. *ACL 2026*.

[c.10] Fanxiao Li, **Jiaying Wu**, Tingchao Fu, Dayang Li, Herun Wan, Wei Zhou, Min-Yen Kan. What's Left Unsaid? Detecting and Correcting Misleading Omissions in Multimodal News Previews. *ACL 2026*.

[c.11] Xiangzheng Kong, Minnan Luo, Wenya Wang, **Jiaying Wu**, Zhi Zeng, Guang Dai. From What Is Said to Why It Is Framed: Intent-Aware News Video Understanding. *ACL 2026 Findings*.

[c.12] **Jiaying Wu**, Fanxiao Li, Zihang Fu, Min-Yen Kan, Bryan Hooi. Seeing Through Deception: Uncovering Misleading Creator Intent in Multimodal News with Vision-Language Models. *ICLR 2026*.

[c.13] Zhi Zeng*, Yifei Yang*, **Jiaying Wu**†, Xulang Zhang, Xiangzheng Kong, Herun Wan, Zihan Ma, Minnan Luo†. From Manipulation to Mistrust: Explaining Diverse Micro-Video Misinformation for Robust Debunking in the Wild. *WWW 2026*.

[c.14] Fanxiao Li, **Jiaying Wu**†, Tingchao Fu, Yunyun Dong, Bingbing Song, Wei Zhou†. Drifting Away from Truth: GenAI-Driven News Diversity Challenges LVLM-Based Misinformation Detection. *AAAI 2026*.

[c.15] Zhengjia Wang, Danding Wang, Qiang Sheng, **Jiaying Wu**, Juan Cao. Reasoning About the Unsaid: Misinformation Detection with Omission-Aware Graph Inference. *AAAI 2026 (Oral)*.

[c.16] Nuo Chen, Andre Lin HuiKai, **Jiaying Wu**, Junyi Hou, Zining Zhang, Qian Wang, Xidong Wang, Bingsheng He. XtraGPT: Context-Aware and Controllable Academic Paper Revision via Human-AI Collaboration. *ACL 2026*.

[c.17] Ziyang Zhang, Yanxia Qin, Ram Pravin Kumar Muthuramalingam, Yixi Ding, Wei Heng Chng, Ru-Chun Dai, **Jiaying Wu**, Chen-Hua Yeow, Min-Yen Kan, Giorgia Pastorin. Quantitative Computational Validation of Nanoscale Interactions between Drug Molecules and Diabetic Wound-Related Proteins for Drug Discovery. *ACS Nano Medicine, 2026*.

[c.18] Yue Liu*, **Jiaying Wu***, Yufei He*, Ruihan Gong*, Jun Xia, Liang Li, Hongcheng Gao, Hongyu Chen, Baolong Bi, Jiaheng Zhang, Zhiqi Huang, Bryan Hooi, Stan Z. Li, Keqin Li. Efficient Inference for Large Reasoning Models: A Survey. *IEEE T-PAMI, 2026*.

[c.19] Zhiyuan Hu*, Yucheng Wang*, Yufei He*, **Jiaying Wu**, Yilun Zhao, See-Kiong Ng, Cynthia Breazeal, Anh Tuan Luu, Hae Won Park, Bryan Hooi. Rewarding the Rare: Uniqueness-Aware RL for Creative Problem Solving in LLMs. *ACL 2026 Findings*.

[c.20] Herun Wan, **Jiaying Wu**, Minnan Luo, Zhi Zeng, Zhixiong Su. Truth over Tricks: Measuring and Mitigating Shortcut Learning in Misinformation Detection. *NeurIPS 2025*.

[c.21] Zhi Zeng, **Jiaying Wu**, Minnan Luo, Xiangzheng Kong, Zihan Ma, Guang Dai, Qinghua Zheng. Understand, Refine and Summarize: Multi-Granularity Knowledge Progressive Enhancement Learning for Fake News Video Detection. *ACM Multimedia 2025*.

[c.22] Zhi Zeng, **Jiaying Wu**, Minnan Luo, Herun Wan, Xiangzheng Kong, Zihan Ma, Guang Dai, Qinghua Zheng. IMOL: Incomplete-Modality-Tolerant Learning for Multi-Domain Fake News Video Detection. *ACL 2025 (Oral)*.

[c.23] Fanxiao Li, **Jiaying Wu**, Canyuan He, Wei Zhou. CMIE: Combining MLLM Insights with External Evidence for Explainable Out-of-Context Misinformation Detection. *ACL 2025 Findings*.

[c.24] Yifei Xu, **Jiaying Wu**†, Herun Wan, Yang Li, Zhen Hou, Min-Yen Kan†. Forecasting the Buzz: Enriching Hashtag Popularity Prediction with LLM Reasoning. *CIKM 2025 (Short)*.

[c.25] Hui Chen, Miao Xiong, Yujie Lu, Wei Han, Ailin Deng, Yufei He, **Jiaying Wu**, Yibo Li, Yue Liu, Bryan Hooi. MLR-Bench: Evaluating AI Agents on Open-Ended Machine Learning Research. *NeurIPS 2025 (Datasets and Benchmarks Track)*.

[c.26] Yibo Li, Miao Xiong, **Jiaying Wu**, Bryan Hooi. ConfTuner: Training Large Language Models to Express Their Confidence Verbally. *NeurIPS 2025*.

[c.27] **Jiaying Wu**, Jiafeng Guo, Bryan Hooi. Fake News in Sheep's Clothing: Robust Fake News Detection Against LLM-Empowered Style Attacks. *KDD 2024 (KDD 2024's Most Influential Paper #7)*.

[c.28] Jianqing Xu, Shen Li, **Jiaying Wu**, Miao Xiong, Ailin Deng, Jiazhen Ji, Yuge Huang, Guodong Mu, Wenjie Feng, Shouhong Ding, Bryan Hooi. ID3: Identity-Preserving-yet-Diversified Diffusion Models for Synthetic Face Recognition. *NeurIPS 2024*.

[c.29] **Jiaying Wu**, Shen Li, Ailin Deng, Miao Xiong, Bryan Hooi. Prompt-and-Align: Prompt-Based Social Alignment for Few-Shot Fake News Detection. *CIKM 2023*.

[c.30] **Jiaying Wu**, Bryan Hooi. DECOR: Degree-Corrected Social Graph Refinement for Fake News Detection. *KDD 2023*.

[c.31] Miao Xiong, Ailin Deng, Pang Wei Koh, **Jiaying Wu**, Shen Li, Jianqing Xu, Bryan Hooi. Proximity-Informed Calibration for Deep Neural Networks. *NeurIPS 2023 (Spotlight)*.

[c.32] Jianqing Xu, Shen Li, Ailin Deng, Miao Xiong, **Jiaying Wu**, Jiaxiang Wu, Shouhong Ding, Bryan Hooi. Probabilistic Knowledge Distillation of Face Ensembles. *CVPR 2023*.

[c.33] **Jiaying Wu**, Bryan Hooi. Probing Spurious Correlations in Popular Event-Based Rumor Detection Benchmarks. *ECML-PKDD 2022*.

Workshop Publications

[w.1] Nuo Chen, Yicheng Tong, **Jiaying Wu**, Minh Duc Duong, Qian Wang, Qingyun Zou, Bryan Hooi, Bingsheng He. Beyond Brainstorming: What Drives High-Quality Scientific Ideas? Lessons from Multi-Agent Collaboration. *AAAI 2026 Workshop on AI for Scientific Research*.

[w.2] Yixi Ding, **Jiaying Wu**†, Tongyao Zhu, Yanxia Qin, Qian Liu, Min-Yen Kan†. CCSBench: Evaluating Compositional Controllability in LLMs for Scientific Document Summarization. *KDD 2025 Workshop on Large Language Models for Scientific and Societal Advances (Oral)*.

[w.3] Sahajpreet Singh, **Jiaying Wu**, Svetlana Churina, Kokil Jaidka. On the Limitations of LLM-Synthesized Social Media Misinformation Moderation. *ICLR 2025 ICBINB Workshop*.

[w.4] **Jiaying Wu**, Bryan Hooi. Chain-of-Timeline: Enhancing LLM Zero-Shot Temporal Reasoning with SQL-Style Timeline Formalization. *ICLR 2025 Workshop on Reasoning and Planning for Large Language Models*.

Honors & Awards

- Gold Reviewer, ICML 2026 — 2026
- Research Achievement Award × 2 (AY 22/23, AY 23/24), NUS School of Computing — 2023 – 2024
- Research Incentive Award, NUS School of Computing — 2023
- NUS Research Scholarship — 2020 – 2024
- Student Award of Merit, BFSU — 2019
- Undergraduate Academic Scholarship × 3 (AY 16/17; 17/18; 18/19), BFSU — 2016 – 2019
- Special Prize (9th / 40,000+ contestants), National English Competition for College Students, China — 2019

Invited Talks

What Should AI Understand About Us? Social Intelligence for Reliable Decision-Making

- Shenzhen University of Advanced Technology — 09/2026
- IR Group at Tsukuba University — 09/2026

The Facade of Truth: Uncovering and Mitigating LLM Susceptibility to Deceptive Evidence

- Technical Sharing Session, Infocomm Media Development Authority (IMDA) — 09/2026

Creator Intent, Omission, and AI-Augmented Correction in Digital Information Resilience

- NUS Centre for Trusted Internet and Community (CTIC) — 04/2026

From Patterns to Purpose: Towards Generalizable and Intent-Aware Misinformation Governance

- State Key Laboratory of AI Safety, Institute of Computing Technology, CAS — 06/2025
- Media Synthesis and Forensics Lab, Institute of Computing Technology, CAS — 06/2025

LLMs for Social Simulation: Progress, Opportunities and Challenges

- NUS Centre for Trusted Internet Community (CTIC) 09/2024

Robust Fake News Detection Against LLM-Empowered Style Attacks

- Key Lab of Network Data Science and Technology, Institute of Computing Technology, CAS 11/2023

Degree-Corrected Social Graph Refinement for Fake News Detection

- Graph Data Mining and Machine Learning Lab (GAMMA Lab), BUPT 11/2023

Teaching

CS4248: Natural Language Processing, NUS Spring 2022

- Teaching Assistant to Min-Yen Kan.
- Duties: led tutorials, advised on course projects, and designed exam problems.

Professional Service

Workshop / Event Organizer

- Organizer, CIVIC-AI 2026: Workshop on Collaborative Intelligence and the Future of Work. [🌐] [whitepaper]
- Organizer, ICLR 2025 Workshop on Reasoning and Planning for Large Language Models. [🌐]
- General Chair, The 2024 Singapore Symposium on Natural Language Processing (SSNLP 2024). [🌐]

Area Chair

- ACL Rolling Review (2026 –), CIKM 2025–2026, NeurIPS 2026, ICLR 2027.

Reviewer

- Conference Reviewer: NeurIPS, ICML, ICLR, ACL ARR, WebConf, KDD, CIKM, CVPR, ACM MM, AISTATS.
- Journal Reviewer: IEEE TKDE, IEEE T-IFS, ACM TIST.

Skills

Technical

- Advanced: Python (Transformers, PyTorch). Experience with: C, Java, R, Matlab, C++, JSP, HTML5/CSS.

Languages

- CATTI-Certified Translator and Interpreter (Chinese ⇔ English, qualified for professional service).

Previous Research Experience

Institute of Computing Technology, CAS Beijing, China
Research Intern advised by Jiafeng Guo 08/2019 – 02/2020

- Project: Neural Question Generation with Controllable Question Words.

National University of Singapore Singapore
Summer Research Student advised by Leong Hon Wai 07/2018 – 08/2018

- Project: Community Discovery in Fashion Runway Photos.

Selected Translated Works

Translator, *The Beast of Buckingham Palace* by David Walliams 2025

- English to Chinese; Jieli Publishing House, China. ISBN: 9787544890335.

Translator, *Storm Boy* by Colin Thiele 2023

- English to Chinese; Jieli Publishing House, China. ISBN: 9787544879859.

Translator, *Make Your Own Museum* by the Victoria and Albert Museum (V&A) 2018

- English to Chinese; Hunan Juvenile & Children's Publishing House, China. ISBN: 9787556247783.

Translator, *How to Fly Like An Elephant* by Kyoko Nemoto 2018

- English to Chinese; Hunan Juvenile & Children's Publishing House, China. ISBN: 9787556245536.

Translator, *Thomas & Friends* Bilingual Picture Books Series 2018

- English to Chinese; Hunan Juvenile & Children's Publishing House, China. ISBN: 9787556242191.

Community Engagement

Volunteer Interpreter, "S.T. Lee Cup" Chinese Youth Chess Tournament 2019

- Interpreter for Mr. Leontxo García, Senior Adviser at Chess in Education Commission, FIDE.

Student Director, BFSU Center for Student Learning and Development 2018 – 2019

- Coordinated student-led programs, events, and invited talks.

Volunteer Camp Counselor, 2017 Soong Ching Ling International Summer Camp (SCLISC) 2017

- Organized camp activities; interpreter for campers from English-speaking countries.