WENJIE WANG

School of Software Engineering, Sichuan University, Chengdu, China
jackwwj619@gmail.com ◇ Jackwwj619 ◇ jackwwj619.github.io

RESEARCH INTERESTS

- **LLM Post-training & Evaluation** — aligning and evaluating LLMs through synthetic data, personalization, and interactive assessment.
- **Trustworthy LLM Reasoning** — identifying and mitigating emergent failure modes in LLM reasoning, from social intelligence risks to conformity cascades.
- **Multi-Agent Systems** — exploring the dynamics, applications, and safety of LLM-based multi-agent collaboration and interaction.

EDUCATION

Sichuan University Chengdu, China
B.S. in Software Engineering, GPA: 3.91/4.0 Sep 2022 – Jun 2027 (expected)

PUBLICATIONS

(* denotes equal contribution / co-first authorship.)

[1] Yu Jiang*, **Wenjie Wang***, Yue Huang*, Yanbo Wang, Zhenhong Zhou, Xiuying Chen, Yang Liu, Pin-Yu Chen, Wei Wang, Xiangliang Zhang. "*RiskLab: A Controlled Toolkit for Probing Emergent Risks in LLM-Based Multi-Agent Systems*." **ACL 2026 (System Demonstrations).** [paper] [code]

[2] Yue Huang*, Yu Jiang*, **Wenjie Wang***, Haomin Zhuang, Xiaonan Luo, Yuchen Ma, Zhangchen Xu, Zichen Chen, Nuno Moniz, Zinan Lin, Pin-Yu Chen, Nitesh V. Chawla, Nouha Dziri, Huan Sun, Xiangliang Zhang. "*Emergent Social Intelligence Risks in Generative Multi-Agent Systems*." **arXiv:2603.27771**, 2026. [paper]

[3] Yuchen Ma, Yue Huang, **Wenjie Wang**, Xiaonan Luo, Xiangliang Zhang, Stefan Feuerriegel. "*Synthetic Interaction Data for Scalable Personalization in Large Language Models*." **KDD D&B 2026.** [paper]

[4] Yue Huang, Yuchen Ma, Jiayi Ye, **Wenjie Wang**, Zipeng Ling, Xingjian Hu, Yuexing Hao, Zichen Chen, Zhangchen Xu, Yunhong He, Zhengqing Yuan, Yujun Zhou, Kehan Guo, Chaoran Chen, Toby Jia-Jun Li, Stefan Feuerriegel, Xiangliang Zhang. "*NARRA-Gym for Evaluating Interactive Narrative Agents*." **arXiv:2605.08503**, 2026. [paper]

RESEARCH EXPERIENCE

Emergent Social Intelligence Risks in Multi-Agent Systems Aug 2025 – Jan 2026
Core Contributor

- Systematically defined a taxonomy of emergent social intelligence risks in generative multi-agent systems, covering collusion-like coordination, conformity cascades, and role-overstepping redundant execution.
- Designed controlled experiments across multiple topologies (hierarchical, peer-to-peer, hybrid) and interaction protocols, with controlled variables including agent count, communication rounds, and role assignment strategies.
- Demonstrated that prompt-level safety constraints and single-agent alignment are insufficient against emergent group risks; proposed mechanism-level governance (anti-collusion design, fairness enforcement, incentive-compatible reporting).

Synthetic Interaction Data for Scalable Personalization in LLMs Feb 2025 – Feb 2026
Core Contributor

- Co-developed PersonaGym, a high-fidelity synthetic data generation framework using an agentic LLM system to simulate realistic user preference behaviors and semantic-aware noise in multi-turn personalized interaction trajectories.
- Generated PersonaAtlas, a large-scale dataset covering diverse user personas, preference intensities, and noise patterns, providing a data foundation for downstream personalization training.
- Verified framework robustness and generalization under noisy and sparse preference signals, achieving superior personalization quality and task performance over baseline methods.

AWARDS & HONORS

National Scholarship, Ministry of Education of China — Oct 2025
Awarded to approximately top 0.2% of undergraduates nationwide.

Excellent Completion, National Undergraduate Innovation Training Program (National-level) — May 2025
Ministry of Education of China; rated "Excellent" among national-level undergraduate research projects.

Mathematical Contest in Modeling (MCM) — Finalist, COMAP — May 2025
Top $\approx$ 2% of 27,000+ teams worldwide.

National Scholarship, Ministry of Education of China — Oct 2024
Awarded to approximately top 0.2% of undergraduates nationwide.