RUI LI

⌂ O-L1RU1.github.io ☎ (+86) 16511106999 ✉ o_l1ru1@stu.pku.edu.cn

Education

Peking University
Master in School of Software and Microelectronics, Artificial Intelligence Sep. 2023 - Present
Advisor: Prof. Zhifang Sui *Beijing, China*

ShanDong University **GPA: 3.89/4.0**
B.S. in School of Computer Science and Technology, Artificial Intelligence Sep. 2019 – Jul. 2023
Advisor: Prof. Qiong Zeng *Shandong, China*

Publications

- **How Far are LLMs from Being Our Digital Twins? A Benchmark for Persona-Based Behavior Chain Simulation**
  Rui Li, Heming Xia, Xinfeng Yuan, Qingxiu Dong, Lei Sha, Wenjie Li, Zhifang Sui
  Proceedings of the 2025 Conference on Association for Computational Linguistics ***ACL 2025 findings***.

- **Towards Harmonized Uncertainty Estimation for Large Language Models**
  Rui Li, Jing Long, Muge Qi, Heming Xia, Lei Sha, Peiyi Wang, Zhifang Sui
  Proceedings of the 2025 Conference on Association for Computational Linguistics ***ACL 2025 (oral)***.

- **Be a Multitude to Itself: A Prompt Evolution Framework for Red Teaming**
  Rui Li, Peiyi Wang, Jingyuan Ma, Di Zhang, Zhifang Sui, Lei Sha
  Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. ***EMNLP 2024 findings***.

- **SuperGPQA: Scaling LLM Evaluation across 285 Graduate Disciplines**
  M-A-P (Multimodal Art Projection), **Core Contributor**
  The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track. ***NeurIPS 2025***

- **Layer-Aware Representation Filtering: Purifying Finetuning Data to Preserve LLM Safety Alignment**
  Hao Li, Lijun Li, Zhenghao Lu, Xianyi Wei, ***Rui Li***, Jing Shao, Lei Sha
  Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. ***EMNLP 2025***.

- **Beyond Single Frames: Can LMMs Comprehend Implicit Narratives in Comic Strip?**
  Xiaochen Wang, Heming Xia, Jialin Song, Longyu Guan, Qingxiu Dong, ***Rui Li***, Yixin Yang, Yifan Pu, Weiyao Luo, Yiru Wang, Xiangdi Meng, Wenjie Li, Zhifang Sui
  Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. ***EMNLP 2025 findings***.

- **A Survey on In-context Learning**
  Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, ***Rui Li***, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, Zhifang Sui
  Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. ***EMNLP 2024***.

- **ShieldLM: Empowering LLMs as Aligned, Customizable and Explainable Safety Detectors**
  Zhexin Zhang, Yida Lu, Jingyuan Ma, Di Zhang, ***Rui Li***, Pei Ke, Hao Sun, Lei Sha, Zhifang Sui, Hongning Wang,

Minlie Huang
Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. ***EMNLP 2024 findings*.**

Under Review & Preprint

* indicates equal contribution

- **LLM-REVal: Can We Trust LLM Reviewers Yet?**
  Rui Li, Jia-Chen Gu, Po-Nien Kung, Heming Xia, Junfeng Liu, Xiangwen Kong, Zhifang Sui, Nanyun Peng
- **Merlin's Whisper: Enabling Efficient Reasoning in LLMs via Black-box Adversarial Prompting**
  Heming Xia, Cunxiao Du, ***Rui Li***, Chak Tou Leong, Yongqi Li, Wenjie Li
- **OS-Catalyst: Advancing Computer-Using Agents Efficiency through Adaptive Action**
  Xinfeng Yuan, Qiushi Sun, Yinghao Chen, ***Rui Li***, Xuetian Chen, Siyu Yuan, Xintao Wang, Zichen Ding, Zonglin Li, Biqing Qi, Deqing Yang
- **SenseJudge: Explicit Preference-Driven Judgment Framework**
  Rui Li*, Junfeng liu*0o, Xiangwen Kong, Zhifang Sui
- **HauntAttack: When Attack Follows Reasoning as a Shadow**
  Jingyuan Ma*, ***Rui Li****, Zheng Li, Junfeng liu, Lei Sha, Zhifang Sui
- **Step-3 is Large yet Affordable: Model-system Co-design for Cost-effective Decoding**
  StepFun, Step-3 technical report, **Contributor**
- **MACG: A Multi-Agent Framework for Thematically Structuring and Generation of Related Work**
  Zhuang Liu, Jian Liu, Chenbin Zhang, ***Rui Li***, Chun Kang, Maolin Wang, Lei Sha
- **Large Language Models Struggle with Unreasonability in Math Problems**
  Jingyuan Ma, Damai Dai, Zihang Yuan, ***Rui Li***, Weilin Luo, Bin Wang, Qun Liu, Lei Sha, Zhifang Sui
- **SCoRE: Benchmarking Long-Chain Reasoning in Commonsense Scenarios**
  Weidong Zhan, Yue Wang, Nan Hu, Liming Xiao, Jingyuan Ma, Yuhang Qin, Zheng Li, Yixin Yang, Sirui Deng, Jinkun Ding, Wenhan Ma, ***Rui Li***, Weilin Luo, Qun Liu, Zhifang Sui
- **Plug-and-Play Training Framework for Preference Optimization**
  Jingyuan Ma, ***Rui Li***, Zheng Li, Lei Sha, Zhifang Sui

Research Experiences

LLM-REVal: Can We Trust LLM Reviewers Yet?
Advisor: Prof. Nanyun Peng, University of California, Los Angeles *April. 2025 – Sep. 2025*

- We propose LLM-REVal (LLM REViewer Re-EValuation) through a multi-round simulation of the academic publication process, incorporating a research agent and a review agent. We conduct human annotations and identify pronounced misalignment between LLM-based reviews and human judgments: (1) LLM reviewers systematically inflate scores for LLM-authored papers, assigning them markedly higher scores than human-authored ones; (2) LLM reviewers persistently underrate human-authored papers with critical statements (e.g., risk, fairness).
- Our analysis reveals two primary biases in LLM reviewers: a linguistic feature bias favoring LLM-generated writing styles, and an aversion toward critical statements (e.g., risk, fairness). These results highlight the risks and equity

concerns posed to human authors and academic research if LLMs are deployed in the peer review cycle without adequate caution.

How Far are LLMs from Being Our Digital Twins? A Benchmark for Persona-Based Behavior Chain Simulation

Advisor: Prof. Zhifang Sui, Peking University *Seb. 2024 – Feb. 2025 Beijing, China*

- To bridge the current research gap in LLM as human digital twins, we propose BehaviorChain, the first benchmark designed to evaluate LLMs' ability to simulate continuous human behaviors. BehaviorChain comprises diverse, high-quality persona-based behavior sequences, encompassing 15,846 distinct behaviors across 1,001 unique personas, extracted from literary corpora using an automated, scalable pipeline.
- Comprehensive evaluations and analysis of ten state-of-the-art LLMs using BehaviorChain revealed that accurately simulating continuous human behaviors remains a significant challenge, even for advanced models like GPT-4o, indicating that the path from LLMs to true digital twins is still long.

Towards Harmonized Uncertainty Estimation for Large Language Models

Advisor: Prof. Zhifang Sui, Peking University *Feb. 2024 – Apr. 2024 Beijing, China*

- Empirical analysis shows that existing uncertainty estimation (UE) methods universally fail to balance the key metrics of Indication, Precision-Recall, and Calibration. We propose CUE (Corrector for Uncertainty Estimation), an orthogonal, lightweight framework that significantly improves UE reliability.
- CUE trains a small auxiliary model (the Corrector) on data reflecting the target LLM's actual performance. By using a weighted integration of the Corrector's scores with scores from existing UE methods, CUE achieves a harmonized balance across all metrics. Experiments show CUE consistently and substantially enhances diverse UE baselines.

Services and Internship

ACL Rolling Review, Reviewer *2024 – 2025*

StepFun AI, Algorithm Intern *Jul. 2024 – Oct. 2025 Beijing, China*

36Kr, Intern Reporter *Jun. 2023 - Sep. 2023 Beijing, China*

Shandong University, Teaching Assistant (Introduction to AI) *Feb. 2023 - Jun. 2023, Shandong, China*

Honors and Awards

- Merit Student, Peking University 2025
- Outstanding Graduate, Shandong University 2023
- Outstanding Undergraduate Graduation Thesis Award and Outstanding Thesis Defense Award 2023
- First Prize in the 14th National College Student Mathematics Competition 2023
- Academic Scholarship and Scientific Innovation Scholarship, Shandong University 2019 - 2022
- First Prize and Application Innovation Award in the Teddy Cup Data Mining Challenge 2022
- Excellent Paper Nomination in the "Shenzhen Cup" Mathematical Modeling Challenge Finals 2022

Technical Skills

Languages: C/C++, Python, Shell, MATLAB, HTML/CSS

Developer Tools: VS Code, PyCharm, Git, Linux, Vim

Minor: Law