RUI LI

Mobile: (+86) 16511106999 ⋄ Email : o_l1ru1@stu.pku.edu.cn

EDUCATION

Peking University **GPA: 3.64/4.0**
Master in School of Software and Microelectronics, Artificial Intelligence Sep. 2023 - Present
Advisor: Prof. Zhifang Sui *Beijing, China*

ShanDong University **GPA: 90.99/100**
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****.*
- **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)****.*
- **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

- **SenseJudge: Explicit Preference-Driven Judgment Framework**
  Rui Li, Junfeng liu, Xiangwen Kong, Zhifang Sui
- **HauntAttack: When Attack Follows Reasoning as a Shadow**
  Jingyuan Ma*, **Rui Li***, Zheng Li, Junfeng liu, Lei Sha, Zhifang Sui
- **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
- **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

- **SuperGPQA: Scaling LLM Evaluation across 285 Graduate Disciplines**
  M-A-P (Multimodal Art Projection), **Core Contributor**
- **KnowLogic: A Benchmark for Commonsense Reasoning via Knowledge-Driven Data Synthesis**
  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
- **Text-driven Palette Generation Method**
  Rui Li, Qiong Zeng
  Outstanding Undergraduate Graduation Thesis of Shandong University.

RESEARCH EXPERIENCES

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 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 a long one.

Towards Harmonized Uncertainty Estimation for Large Language Models
Advisor: Prof. Zhifang Sui, Peking University *Feb. 2024 – Apr. 2024 Beijing, China*

- Identify limitations in current uncertainty estimation methods due to inherent biases in LLMs, such as over-confidence and under-confidence, providing both theoretical proof and empirical evidence.
- Propose an external insight-driven approach that seamlessly integrates with existing uncertainty estimation methods, correcting inversion of uncertainty score rankings caused by LLM biases.

Be a Multitude to Itself: A Prompt Evolution Framework for Red Teaming
Advisor: Prof. Zhifang Sui, Peking University *Apr. 2023 – Feb. 2024 Beijing, China*

- Propose a red teaming prompt evolution framework for LLMs that automatically enhances the quantity and quality of attack prompts, eliminating the need for meticulous prompt crafting.
- Systematically evaluate closed-source and open-source LLMs on various sensitive topics, analyzing them across dimensions such as temporal aspects, scale, and category spans, while providing detailed discussions on variations in pre-generated attack prompts.

Text-driven Palette Generation Method
Advisor: Assoc. Prof. Qiong Zeng, ShanDong University *Jan. 2023 – May. 2023 Shandong, China*

- Build a CGAN-based generative model that integrates joint semantic and color spaces. The generator employs a seq2seq model with a fusion attention mechanism to produce palettes aligned with text semantics, while the discriminator uses fully connected layers to evaluate palette authenticity.
- Dvelop a text-palette dataset for model training. The diversity of generated palettes was quantitatively assessed using color distance, while their rationality and effectiveness were analyzed through CLIP text-image similarity and user experiments.

HONORS AND AWARDS

- **Outstanding Graduate, Shandong University** 2023

- First Prize in the 14th National College Student Mathematics Competition 2023
- Honorable Mention in the First National Advanced Technology Innovation Competition 2023
- Academic Scholarship and Scientific Innovation Scholarship 2019 - 2022
- First Prize and SPSSPRO Application Innovation Award in the Teddy Cup Data Mining Challenge 2022
- Honor Award in the Mathematical Modeling Competition for American College Students 2022
- Excellent Paper Nomination in the "Shenzhen Cup" Mathematical Modeling Challenge Finals 2022

TECHNICAL SKILLS

Languages: C/C++, Python, Java, Shell, MATLAB, HTML/CSS
Developer Tools: VS Code, PyCharm, Git, Linux, Vim
Minor: Law