Jiarui Liu

@ jiaruil5@andrew.cmu.edu | Pittsburgh, PA | +1 971-506-9587

Homepage | Google Scholar | Twitter | LinkedIn | GitHub

Research interests: Responsible AI for social good, LLM reasoning, large language models, multi-modal machine learning, natural language processing.

EDUCATION

Carnegie Mellon University   Pittsburgh, Pennsylvania, US
PhD / Master of Language Technologies (MLT), SCS, LTI; GPA: 4.14/4.33   *Aug. 2023 – Aug. 2028 (Expected)*
- **Advisor:** Mona Diab, department head of LTI at CMU
- **Relevant coursework:** LLM Systems (A+), Neural Code Generation (A+), Advanced NLP (A), Multi-modal ML (A), Quantitative Evaluation

University of Michigan, Ann Arbor   Ann Arbor, Michigan, US
BSE in Computer Science (Dual Degree); GPA: 3.86/4.00   *Sep. 2021 – May 2023*
- **Relevant coursework:** Intro to NLP (A), Science in Deep Learning (A), Operating Systems, Machine Learning (A), Deep Learning in CV (A+), Data Structures and Algorithms (A), Practical Programming in Java (A+), Computer Architecture (A), Web Systems (A-), Mobile App Development (A)

Shanghai Jiao Tong University Joint Institute (UM-SJTU JI)   Shanghai, China
BSE in ECE (Dual Degree);   *Sep. 2019 – Aug. 2023*
- **Relevant coursework:** Programming and Elementary Data Structures, Probability Methods and Statistics, Linear Algebra, Discrete Mathematics, Honor Mathematics

Instituto Tecnológico de Buenos Aires   Buenos Aires, Argentina
Visiting Student   *Jan. 2020 – Feb. 2020*

PUBLICATIONS

∗ indicates equal contribution. † indicates equal supervision.

15. *"Language Model Alignment in Multilingual Trolley Problems"*

    Zhijing Jin∗, Max Kleiman-Weiner∗, Giorgio Piatti∗, Sydney Levine, **Jiarui Liu**, Fernando Gonzalez, Francesco Ortu, András Strausz, Mrinmaya Sachan, Rada Mihalcea, Yejin Choi, Bernhard Schölkopf

    ICLR 2025 and Best Paper Award at NeurIPS Pluralistic Alignment Workshop 2024. [Arxiv]

14. *"Implicit Personalization in Language Models: A Systematic Study"*

    Zhijing Jin∗, Nils Heil∗, **Jiarui Liu**∗, Shehzaad Dhuliawala∗, Yahang Qi∗, Bernhard Schölkopf, Rada Mihalcea, Mrinmaya Sachan

    EMNLP 2024 Findings. [Arxiv] [Code] [Video]

13. *"Synatra: Turning indirect knowledge into direct demonstrations for digital agents at scale"*

    Tianyue Ou, Frank F Xu, Aman Madaan, **Jiarui Liu**, Robert Lo, Abishek Sridhar, Sudipta Sengupta, Dan Roth, Graham Neubig, Shuyan Zhou

    NeurIPS 2024. [Arxiv]

12. *"Inducing Elasticity in Foundation Models: Post-Training Techniques for Adaptable Inference"*

    Aashiq Muhamed, **Jiarui Liu**, Mona Diab, Virginia Smith

    NeurIPS ENLSP Workshop 2024.

11. *"Automatic Generation of Model and Data Cards: A Step Towards Responsible AI"*

    Jiarui Liu, Wenkai Li, Zhijing Jin, Mona Diab.

    NAACL 2024 Oral. [ACL Anthology] [Arxiv] [Video] [Code]

10. *"Analyzing the Role of Semantic Representations in the Era of Large Language Models"*

Zhijing Jin*, Yuen Chen*, Fernando Gonzalez*, **Jiarui Liu**, Jiayi Zhang, Julian Michael, Bernhard Schölkopf, Mona Diab

NAACL 2024. [ACL Anthology] [Arxiv]

9. *"Can Large Language Models Infer Causation from Correlation?"*
Zhijing Jin*, **Jiarui Liu***, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, Bernhard Schölkopf

ICLR 2024. [Arxiv] [Code]

8. *"Bias Amplification Enhances Minority Group Performance"*
Gaotang Li*, **Jiarui Liu***, and Wei Hu.

TMLR 2023. [Arxiv] [Video]

Papers Under Review & Preprints

7. *"Towards Global AI Inclusivity: A Large-Scale Multilingual Terminology Dataset (GIST)"*
Jiarui Liu*, Iman Ouzzani*, Wenkai Li*, Lechen Zhang, Tianyue Ou, Houda Bouamor, Zhijing Jin, Mona Diab

ARR 2025 Under Review. [Arxiv]

6. *"TACO: Taming Multimodal Hallucinations with Contrastive-Aware Self-Confidence Calibration"*
Jiarui Liu, Renato Negrinho, Manuel Mager, Shang Chao, Ren Pang, Neha Anna John, Yassine Benajiba

ARR 2025 Under Review.

5. *"BIG5-CHAT: Shaping LLM Personalities Through Training on Human-Grounded Data"*
Wenkai Li*, **Jiarui Liu***, Andy Liu, Xuhui Zhou, Mona T. Diab, Maarten Sap

ARR 2025 Under Review. [Arxiv] [Code]

4. *"Uncovering and Understanding Social Media Censorship across Countries"*
Neemesh Yadav*, **Jiarui Liu***, Francesco Ortu, Zhijing Jin, Rada Mihalcea

ARR 2025 Under Review.

3. *"Chumor 2.0: Towards Benchmarking Chinese Humor Understanding"*
Ruiqi He, Yushu He, Longju Bai, **Jiarui Liu**, Zhenjie Sun, Zenghao Tang, He Wang, Hanchen Xia, Rada Mihalcea, Naihao Deng

Preprint 2024. [Arxiv]

2. *"Chumor 1.0: A Truly Funny and Challenging Chinese Humor Understanding Dataset from Ruo Zhi Ba"*
Ruiqi He, Yushu He, Longju Bai, **Jiarui Liu**, Zhenjie Sun, Zenghao Tang, He Wang, Hanchen Xia, Naihao Deng

Preprint 2024. [Arxiv]

1. *"Voices of Her: Analyzing Gender Differences in the AI Publication World"*
Yiwen Ding*, **Jiarui Liu***, Zhiheng Lyu*, Kun Zhang, Bernhard Schoelkopf, Zhijing Jin[†], and Rada Mihalcea[†].

Preprint 2023. [Arxiv] [Code]

Work Experience

Amazon AWS   New York, US
Applied Scientist Intern   *Jun. 2024 – Aug. 2024*
Host: Yassine Benajiba, Renato Negrinho, Manuel Mager, Neha Anna John

- Developed a sampling-based confidence calibration approach to mitigate object hallucination in Visual Question Answering (VQA) tasks.
- Introduced a novel atomic fact verification and refinement pipeline, reaching SOTA on 5 hallucination

benchmarks and 2 multimodal models.
- Conducted experiments across both black-box and gray-box access to model logits and utilized visual contrastive decoding to further enhance calibration accuracy.

WarpEngine   Shanghai, China
Natural Language Processing R&D Intern, Host: Hang Chu, Ming Liang   *May 2023 – Aug. 2023*
- Constructed 14B language models with customized personalities by adopting different quantization, acceleration, and parameter efficient fine tuning methods (LoRA, P-Tuning).
- Built a pipeline for generating and cleaning customized dialogue data.
- Developed a comprehensive framework for evaluating open-domain dialogue capabilities of large language models.

Selected Research Experience

Automated Model Card Generation & Translation @R3LIT Lab   Advised by Mona Diab
Graduate Research Assistant   *Carnegie Mellon University, Sept. 2023 – Now*
- Constructed a dataset of 10,000 model cards with direct links to corresponding papers and GitHubs.
- Developed a hierarchical retrieve-and-generate system to automatically generate model and dataset cards for Hugging Face.
- Evaluated the proposed pipeline using standard faithfulness metrics, GPT-based metrics, and human evaluation, demonstrating its effectiveness and comprehensiveness.
- Collected English AI terminologies at scale and translated them into Arabic, Chinese, French, Japanese, and Russian through a combination of LLM-based and human validation, exploring its integration and applications in machine translation.
- Led projects, resulting in an Oral paper accepted at NAACL and another under review.

Human-Grounded LLM Personality Induction   Advised by Mona Diab and Maarten Sap
Graduate Research Assistant   *Carnegie Mellon University, May 2024 – Oct. 2024*
- Developed BIG5-CHAT, a large-scale dataset of 100,000 dialogues to ground LLMs in realistic human personality expression through supervised fine-tuning and direct preference optimization.
- Demonstrated that training-based personality alignment methods outperform prompt-based approaches in assessments such as BFI and IPIP-NEO, with findings highlighting trait-based impacts on reasoning tasks.
- Led the project, resulting in a paper currently under review for ARR 2025.

NLP for Social Good @LIT Group   Advised by Rada Mihalcea and Zhijing Jin
Undergraduate Research Assistant   *University of Michigan, Apr. 2022 – Now*
- Constructed the 78k AI SCHOLAR dataset with 20+ features such as gender, affiliation, and domains of specialization, and conducted comprehensive statistical analyses on subgroups of gender, academic age, citation trends, etc.
- Designed an AI Scholar Toolbox for Twitter account look-up, and achieved 80% F1-score.
- Built a Python package for the comparison of two corpora including linguistic differences and classification error analysis of transformer-based models.
- Resulted in three papers accepted and three papers under review.

Exploring Causality in LLMs @LIT Group   Advised by Rada Mihalcea and Zhijing Jin
Undergraduate Research Assistant   *University of Michigan, Apr. 2022 – Apr. 2023*
- Fine-tuned and prompted 12 BERT-based and GPT-based models on the CORR2CAUSE dataset.
- Led experiments and conducted a robustness analysis of the models' causal discovery capabilities, focusing on paraphrasing and variable refactorization.
- Resulted in a paper accepted at ICLR 2024.

Enhancing Worst-Group Robustness @Wei Hu's Group   Advised by Wei Hu
Undergraduate Research Assistant   *University of Michigan, Jul. 2022 – Dec. 2022*
- Developed a two-stage algorithm to enhance worst-group accuracy using a trainable instance-wise auxiliary variable.

- Designed and conducted experiments, achieving state-of-the-art performance on benchmarks including Waterbirds, CelebA, MultiNLI, and CivilComments.
- Led the project, culminating in a published journal paper in TMLR.

Professional Service

Conference Reviewer: ICLR 2025, CHI 2025
Workshop Reviewer: NeurIPS ENLSP 2024

Advising & Mentoring

Iman Ouzzani, 2024, BS at CMU Qatar

Skills

Programming: Python, C/C++, R, Java, Javascript, Go

Framework: PyTorch, Tensorflow, Transformers, Accelerate, DeepSpeed, BitsandBytes, NLTK, Scikit-Learn, XGBoost

Languages: Chinese (Native), English (Proficient, TOEFL 107 (S25 W30 R25 L27))

Awards & Achievements

Future Technology Taihu Scholarship Shanghai Jiao Tong University, Jul. 2023. $1500.

James B. Angell Scholar University of Michigan, Mar. 2023.

Dean's Honor List: University of Michigan, 2021, 2022.

University Honors: University of Michigan, 2021, 2022.

Finalist Winner (Top 2%) in the Interdisciplinary Contest in Modelling (ICM): Collaborated with 2 team members to construct methods of evaluating food systems, come up with suited prediction algorithms, and build visualization models. UM-SJTU JI, Feb. 2021.

Undergraduate Excellent Scholarship: UM-SJTU JI, Jan. 2021.