# Xuran LI

(0406 821 924) · lixuran93@gmail.com · Sydney, Australia

## Education

**University of New South Wales (UNSW Sydney)** *Aug 2025 – Present*
Doctor of Philosophy (PhD) in Computer Science and Engineering
Collaborative Human-Centric AI Systems (CRUISE) Lab

**University of Chinese Academy of Sciences (UCAS)** *Sep 2019 – Jun 2022*
Master of Engineering in Software Engineering
State Key Laboratory of Computer Science, Institute of Software

**Chang'an University** *Sep 2012 – Jun 2016*
Bachelor of Engineering in Network Engineering

## Research Interests

**Trustworthy Machine Learning:** AI for Social Impact; Robustness, Fairness, and Privacy; Adversarial Learning; Reliable AI Systems.
**Large Language Model Personalization:** User-Centered Personalization and Generalization, Semantic Constraint Personalization.

## Publications

1. **Xuran Li**, Peng Wu, Jing Su. *Accurate Fairness: Improving Individual Fairness without Trading Accuracy.* Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI 2023 | CCF A |Core A*).
2. **Xuran Li**, Hao Xue, Peng Wu, Xingjun Ma, Zhen Zhang, Huaming Chen, Flora D. Salim. *Perturbation Effects on Robustness and Individual Fairness.* The 2026 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2026 | CCF A | Core A*).
3. Jing Su, Zhen Zhang, Peng Wu, **Xuran Li**, Jian Zhang. *Adversarial Input Detection Based on Critical Transformation Robustness.* Proceedings of the 33rd International Symposium on Software Reliability Engineering (ISSRE 2022 | CCF B | Core A).
4. Zhen Zhang, Peng Wu, Yuting Yang, **Xuran Li**. *Interpreting Deep Neural Networks via Relative Activation-Deactivation Abstractions.* ACM Transactions on Software Engineering and Methodology (TOSEM 2025 | CCF A | Core A*).

## Manuscripts Under Review / Preprints

1. **Xuran Li**, Guanqin Zhang, Imran Razzak, Hakim Hacid, Eleanna Kafeza, Hao Xue, Flora D. Salim. *Evaluating LLM Personalization via Semantic Constraint Verification*(arXiv preprint 2026).

## Research Experience

**Master Student (Trustworthy Machine Learning)**
State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences (UCAS)
Beijing, China *Sep 2019 – Sep 2024*

- Conducted research under Prof. Peng Wu and Prof. Jian Zhang on trustworthy machine learning.
- Proposed **Accurate Fairness** (Li et al., AAAI 2023) to reconcile prediction accuracy with individual fairness.
- Researched adversarial input detection (Su et al., ISSRE 2022; Zhang et al., TOSEM 2025), leveraging differences between clean and adversarial inputs.
- Developed **RIFair** (Li et al., KDD 2026) to evaluate robustness and fairness under *semantic-preserving* (truth-condition-preserving) perturbations, ensuring predictions remain correct and invariant across semantically equivalent individuals.

**Research Assistant**
Research Center for Brain-inspired Intelligence, Institute of Automation, Chinese Academy of Sciences (UCAS)
Beijing, China *Jun 2022 – Nov 2022*

- Worked under Prof. Yi Zeng, Deputy Director of the Research Center for Brain-inspired Intelligence.
- Developed the **Accurate Fairness Evaluator** using Python, NumPy, TensorFlow, and Keras on the AI Governance Online platform: https://www.aigovernance.online/fairnessevaluator/.

**Research Assistant**
Fudan Vision and Learning (FVL) Lab, Fudan University
Shanghai, China *Oct 2023 – Aug 2024*

- Worked under Prof. Xingjun Ma on robust individual fairness (RIFair) (Li et al., arXiv 2025).
- Proposed the **RIFair** method to expose false or biased adversarial defects in DNNs, analyzing impacts on both prediction accuracy and individual fairness.

**PhD Student (LLM Personalization & Trustworthy Machine Learning)**
Collaborative Human-Centric AI Systems (CRUISE) Lab, University of New South Wales (UNSW)
Sydney, Australia *Sep 2024 – Present*

- Conducted research under Prof. Flora Salim (Deputy Director, UNSW AI Institute) and Prof. Hao Xue (Professor, The Hong Kong University of Science and Technology (Guangzhou), HKUST) on LLM personalization assessment and improvement.
- Developed robust individual fairness evaluation frameworks (**RIFair**) to jointly characterize and assess susceptibility of prediction accuracy and individual fairness under semantic-preserving perturbations.
- Research on context-aware personalization and generalization under semantic constraints in large language models.