Yi Yu (Eve), Ph. D.

Current Position: Young Researcher @ Shanghai Artificial Intelligence Laboratory
Email: evening4g@gmail.com · LinkedIn · Personal Homepage: eveyuyi.github.io

SUMMARY

Experienced researcher focusing on **responsible AI**, including **MLLM Safety Alignment**, **AI security**, and **AI-generated Content Detection**. Proven track record of publications (25+) developing innovative AI security solutions with an intersection of **smart cities**. Demonstrated excellence through **interdisciplinary collaborations** and high-quality publications in top-tier AI conferences and journals.

RESEARCH INTERESTS

LLM Security and Safety Alignment, **Trustworthy AI**, **AI Agent**, **AI-generated Content Detection**, **LLM Watermarking**, **Reinforcement Learning**, **Multi-Agent Systems and Evaluation**, **Smart Cities**

EDUCATION

2017-2022	**Zhejiang University**	Doctor of Philosophy	Transp. Eng.	Advisor:Prof. Dianhai Wang
2020-2022	**Imperial College London**	Visiting Scholar	Advisor: Prof. Washington Ochieng, FREng	
2013-2017	**Zhejiang University**	Bachelor of Science	Major in Civil Engineering	GPA 3.72/4.0
			Minor in Law	GPA 3.78/4.0

*Zhejiang University is ranked 51 in US News Ranking, and ranked 44 in QS Rankings

WORK AND INTERNSHIP EXPERIENCE

2022.07-Present **Shanghai AI Lab** AI Safety Center **Young Researcher**

- Pioneer reinforcement learning-based post-training methods for LLM alignment, applying GRPO to enhance models' persuasion resistance while maintaining helpfulness without compromising general capabilities.
- Lead research on AI security and trustworthiness, drafting AI responsible scaling policies, developing novel LLM post-training alignment methods against multi-turn jailbreaks, balancing security and usability.
- Develop advanced frameworks for AI-generated content detection, establishing real-world benchmarks (e.g., EvoBench).
- Design and implement multi-agent simulation platforms to evaluate the safety and trustworthiness of MLLMs
- Manage multidisciplinary teams and facilitated developing an LLM-based traffic data trading platform.

2020.06-2020.10 **Didi Chuxing Tech Co.** Safety Product Department **Algorithm Intern**

- Developed machine learning algorithms for anomaly detection and risk assessment in ride-sharing systems.
- Proposed anti-cheating algorithms and secured four national patents.

SELECTED PUBLICATIONS AND PRESENTATIONS

Conference Proceedings

1. Jinwei Sun, **Yi Yu***, et al., Backfire-R1: Identifying and Mitigating Persuasion Vulnerabilities in LLM Agents, *ACL 2026* (submitted). [**Reinforcement Learning**, LLM Alignment, Agent Security]
2. Boxuan Zhang*, **Yi Yu***, et al., Dive into the Agent Matrix: A Realistic Evaluation of Self-Replication Risk in LLM Agents, *ICLR 2026* (submitted). [AI Security, LLM Agent]
3. Xiao Yu*, **Yi Yu***, et al., EvoBench: Towards Real-world LLM-Generated Text Detection Benchmarking, *ACL 2025*. [AI Security, Deepfake Detection]
4. Xiaoya Lu, et al., **Yi Yu**, X-Boundary: Establishing Exact Safety Boundary to Shield LLMs from Multi-Turn Jailbreaks without Compromising Usability, *EMNLP 2025*. [AI Safety Alignment]
5. **Yi Yu**, et al., Data on the Move: Traffic-Oriented Data Trading Platform Powered by AI Agent with Common Sense, *IEEE IV 2024*. [Multi-agent, LLM, Smart Cities]
6. **Yi Yu**, et al., SWDPM: A Social Welfare-Optimized Data Pricing Mechanism, *IEEE SMC 2023*.
7. **Yi Yu**, et al., Pursuing Equilibrium of Medical Resources via Data Empowerment in Parallel Healthcare System, *IEEE SMC 2023*. [Optimization, Operations Research]

Journal Articles

1. Danhui Yang, **Yi Yu***, et.al., CycloneGPT: An Iterative Retrieval-Augmented LLM for Expert-Level Cyclone Design, *Applied Energy* (under review).
2. Wang Xuhong, Haoyu Jiang, **Yi Yu**, et al., Building Intelligence Identification System via Large Language Model Watermarking: A Survey and Beyond, *Artificial Intelligence Review*, 2025. [LLM Watermarking]
3. **Yi Yu**, et al. Identifying traffic clusters in urban networks based on graph theory using license plate recognition data. *Physica A: Statistical Mechanics and its Applications*, 2022. [Machine Learning, Smart Cities]
4. **Yi Yu**, et al., Navigating the Data Trading Crossroads: Interdisciplinary Survey, *IEEE TSMC*(submitted).
5. Jiaqi Zeng, **Yi Yu**, et al., Trajectory-as-a-Sequence: A novel travel mode identification framework. *Transportation Research Part C: Emerging Technologies*, 2023. [Deep Learning, Smart Cities]
6. HongSheng Qi, **Yi Yu**, et al., Intersection traffic deadlock formation and its probability: A petri net-based modeling approach. *IET Intelligent Transport Systems*, 2022. [Modelling, Smart Cities]
7. Yanlei Cui, **Yi Yu**, et al., Optimizing Road Network Density Considering Automobile Traffic Efficiency: Theoretical Approach. *Journal of Urban Planning and Development*, 2022. [Smart Cities]

Presentations and Conferences

- **Oral presentation**: "Data on the Move: Transportation-Oriented Data Trading Platform Powered by AI Agent with Common Sense" at 2024 IEEE Intelligent Vehicles Symposium, June 2024, Jeju Island, South Korea
- **Oral presentation**: "SWDPM: A Social Welfare-Optimized Data Pricing Mechanism" at 2023 IEEE International Conference on Systems, Man, and Cybernetics, October 2023, Hawaii, USA.
- **Invited talk**: "Urban Traffic State Monitoring Based on Automatic Number Plate Recognition Data" at 2022 China-UK Technology Summit, December 2022, London, UK.

RESEARCH PROJECTS

AI Generated Contents Detection (National Key R&D Program of China)
Data Economy Research in Smart City (National Key R&D Program of China)
Automatic AI Safety Evaluation System Research (National Key R&D Program of China)
City Brain-ITS: Developing Intelligent Transportation System for Hangzhou (Local Cooperation Project)
Urban traffic intrinsic acquisition and demand structure optimization control based on big data (National Natural Science Foundation of China)
Multi-source Heterogeneous Big Data for Urban Traffic (National Natural Science Foundation of China)
Urban Traffic Structure Control Based on System Dynamics (National Natural Science Foundation of China)

SKILLS

Programming: Python, Shell Scripts, SQL, MATLAB, HTML
Frameworks: PyTorch, TensorFlow, OpenRLHF, VERL, scikit-learn
Expertise: AI Security, AI Safety Alignment, AI-generated Content Detection, Watermarking, Multi-Agent Systems, Large-scale Model Training, MLLM post-training
Software: SUMO, VISSIM, TransCAD, ArcGIS, AutoCAD, LaTeX
Languages: Mandarin (Native), English (Full Professional, TOEFL 103)

ACHIEVEMENTS AND AWARDS

- Published over 25 papers (conference and journal) in high-impact venues covering AI Security, Trustworthy AI, Data-driven Optimization, and Smart City Applications.
- Published/submitted research papers to top-tier conferences including ACL, ICML, and IEEE series.
- Led multiple AI safety projects funded by National Key R&D Programs
- Invited Speaker at international conferences including China-UK Technology Summit, London, 2022.
- Excellent Graduate Student Award, Zhejiang University

SERVICES

Journal Reviewer: IEEE TIV, IEEE TSMC, Applied Sciences, Sustainability
Conference Reviewer: ICML, ICLR, ARR, TRB Annual Meeting, IEEE ITSC, IEEE SMC
Open Source Projects: Contributor to xiaohongshu-mcp (7k Stars)