

Changsheng Sun

Ph.D. Candidate
National University of Singapore

Email: cssun@u.nus.edu

Mobile: +65-8039-5393

Homepage: sunchangsheng.com

EDUCATION

Ph.D. in Computer Science (Supervisor: Prof. Dong Jin Song)

NUS, 2022–2026 (expected)

M.Comp. in Artificial Intelligence

NUS, 2020–2021

B.Eng. in Computer Science

Xidian University, 2015–2019

EXPERIENCE

SoTA Lab, Zhihu Inc.

Singapore

Research Intern, Project Lead (SoTA-LH)

Jun. 2026 – Present

Research Topic: Long-horizon agent evaluation, verifiable user-agent simulation, and verifiable task and trajectory data for post-training. Coding agents, general computer-use agents' benchmarking and RLVR.

Key Result: A new series of LH (Infinite-horizon) benchmarks supported by both real user generated data and formal verifiable data.

National University of Singapore

Singapore

Ph.D. Researcher, Trustworthy AI

Jan. 2022 – Present

Research Topic: LLM safety, adversarial robustness, risk-aware optimization, and reliable/explainable graph learning.

Singtel-NUS Cyber Security R&D Lab

Singapore

Research Intern

Jan. 2020 – Jun. 2020

Research Topic: Deployment-time failure detection for deep neural networks using inter-layer signals.

JD Intelligent Cities Research, JD.com, Inc.

Beijing, China

Algorithm Engineer & Research Intern

Jan. 2018 – Jun. 2018

Research Topic: Ray-based distributed DQN for Spatial-temporal data.

SELECTED PUBLICATIONS

[Google Scholar](#): 600+ citations, h-index: 7. *: corresponding author

- [1] **From Attack Surfaces to Actual Operations: A Survey of Modern LLM Jailbreaks**
Ruikang Zhou, **Changsheng Sun***, Mark Huasong Meng
ACL Findings, 2026.
- [2] **Ignoring Directionality Leads to Compromised Graph Neural Network Explanations**
Changsheng Sun, Xinke Li, Jin Song Dong
IEEE S&P (HMI-SA), 2025.
- [3] **PointCVA: Risk-optimized Outlier Removal for Robust 3D Point Cloud Classification**
Xinke Li, Junchi Lu, Henghui Ding, **Changsheng Sun**, Joey Tianyi Zhou, Chee Yeow Meng
AAAI, 2024.
- [4] **Repairing Failure-inducing Inputs with Input Reflection**
Yan Xiao, Yun Lin, Ivan Beschastnikh, **Changsheng Sun**, David S. Rosenblum, Jin Song Dong
ASE, 2022.
- [5] **Self-Checking Deep Neural Networks in Deployment**
Yan Xiao, Ivan Beschastnikh, David S. Rosenblum, **Changsheng Sun**, Sebastian Elbaum, Yun Lin, Jin Song Dong
ICSE, 2021.
- [6] **Digraph Inception Convolutional Networks**
Zekun Tong, Yuxuan Liang, **Changsheng Sun**, Xinke Li, David S. Rosenblum, Andrew Lim
NeurIPS, 2020.
- [7] **Risk-Aware Robust Graph Network Explanation**
Changsheng Sun, Xinke Li, Jin Song Dong
Manuscript under review, 2026.

ACADEMIC SERVICE

Program Committee / Reviewer:

NeurIPS, ICML, AAAI, WWW, FSE, ASE, ISACE (2022–2026)

Teaching Assistant:

CS5232 Formal Specification & Design Techniques (2024);

CS4218 Software Testing (2023);

CS4211 Formal Methods for Software Engineering (2022)

Awards:

Graduate Student Travel Grants (IEEE S&P 2025, AAAI 2024, ASE 2022);

NUS Research Scholarship (2022–2026)

SoC Tuition Fee Support Award (2026)

Updated: 2026-07