Research Statement

Sun Jun
School of Computing and Information Systems, Singapore Management University
Tel: (65) 6828-1312; Email: junsun@smu.edu.sg
2 (Day) 12 (Month) 2025 (Year)

Background

My research interests center on the application of formal methods to enhance the safety and security of a wide range of systems, with a recent emphasis on fundamental AI models and AI-enabled systems. I am particularly drawn to the potential of systematic and rigorous approaches, grounded in formal reasoning, to bring greater structure and reliability to complex technological landscapes. By fostering a more organized and predictable foundation for these systems, I aspire to contribute to creating a world that is not only more secure but also more conducive to human flourishing and enjoyment.

Research Areas

AI Safety and Security: My research group has been actively engaged in a series of studies focused on multiple critical aspects of AI safety and security, including:

1. Evaluating the safety and security of foundational AI models,
2. Developing systematic methodologies for improving the safety and security of AI systems, and
3. Establishing frameworks for certifying the safety and security of AI models and AI-enabled systems.

Our goal is to advance this line of inquiry in the coming years by developing innovative techniques and tools with tangible impact. These contributions may take the form of theoretical breakthroughs that reshape the community's understanding and approach to AI safety, or practical methodologies that gain widespread adoption in industry. We firmly believe that the importance of addressing AI safety and security cannot be overstated, as it is foundational to the responsible development and deployment of AI technologies.

New Approaches to Software Engineering: In addition to AI safety, my research group is investigating the transformative impact of AI on software engineering practices. Specifically, we are conducting a series of studies to evaluate whether AI technologies could potentially replace human programmers soon. This research is particularly significant given its potential to directly affect millions of software engineers worldwide. By understanding and addressing these changes, we aim to contribute to the evolution of software engineering practices in ways that ensure their relevance and utility in an increasingly AI-driven landscape.

Selected Publications and Outputs

1. Bing Sun, Jun Sun, and Wei Zhao: **Democratic Training Against Universal Adversarial Perturbations**, *ICLR 2025*.
2. Mengdi Zhang, Kai Kiat Goh, Peixin Zhang, Jun Sun, Lin Xin Rose, Hongyu Zhang: **LLMScan: Causal Scan for LLM Misbehavior Detection**, *ICML 2025*.
3. Nay Myat Min, Long H. Pham, Yige Li, Jun Sun: **CROW: Eliminating Backdoors from Large Language Models via Internal Consistency Regularization**, *ICML 2025.*
4. Yedi Zhang, Yufan Cai, Xinyue Zuo, Xiaokun Luan, Kailong Wang, Zhe Hou, Yifan Zhang, Zhiyuan Wei, Meng Sun, Jun Sun, Jing Sun, and Jin Song Dong: **Position: Trustworthy AI Agents Require the Integration of Large Language Models and Formal Methods**, *ICML 2025*.
5. Yige Li, Hanxun Huang, Yunhan Zhao, Xingjun Ma, Jun Sun: **BackdoorLLM: A Comprehensive Benchmark for Backdoor Attacks and Defenses on Large Language Models**, *NeurIPS 2025*.
6. PeiYuan Tang, Xiaodong Zhang, Chunze Yang, Haoran Yuan, Jun Sun, Danfeng Shan, and Zijiang James Yang: **Unleashing the Power of Visual Foundation Models for Generalizable Semantic Segmentation**, *AAAI 2025*.
7. Zongxin Liu, Zhe Zhao, Fu Song, Jun Sun, Pengfei Yang, Xiaowei Huang, and Lijun Zhang: **Training Verification-Friendly Neural Networks via Neuron Behavior Consistency**, *AAAI 2025*.
8. Guoliang Dong, Haoyu Wang, Jun Sun, and Xinyu Wang: **Evaluating and Mitigating Linguistic Discrimination in Large Language Models: Perspectives on Safety Equity and Knowledge Equity**, *IJCAI 2025*.
9. Yufan Cai, Zhe Hou, David Sanan, Xiaokun Luan, Yun Lin, Jun Sun, and Jin Song Dong: **Automated Program Refinement: Guide and Verify Code Large Language Model with Refinement Calculus**, *POPL 2025.*
10. Jinhao Dong, Jun Sun, Wenjie Zhang, Jinsong Dong, and Dan Hao: **ConTested: Consistency-Aided Tested Code Generation with LLM**, *ISSTA 2025*.
11. Bozhi Wu, Chengjie Liu, Zhiming Li, Yushi Cao, Jun Sun, and Shang-Wei Lin: **Enhancing Vulnerability Detection via Inter-procedural Semantic Completion**, *ISSTA 2025*.
12. Jianlei Chi, Xiaotian Wang, Yuhan Huang, Lechen Yu, Di Cui, Jianguo Sun, and Jun Sun: **REACCEPT: Automated Co-evolution of Production and Test Code Based on Dynamic Validation and Large Language Models**, *ISSTA 2025*.
13. Shuang Liu, Chenglin Tian, Jun Sun, Ruifeng Wang, Wei Lu, Yongxin Zhao, Yinxing Xue, Junjie Wang, and Xiaoyong Du: **Semantic Conformance Testing of Relational DBMS**, *VLDB 2025*.
14. Yedi Zhang, Lei Huang, Pengfei Gao, Fu Song, Jun Sun, and Jinsong Song: **Verification of Bit-Flip Attacks against Quantized Neural Networks**, *OOPSLA 2025*.
15. Bo Wang, Chong Chen, Ming Deng, Junjie Chen, Xing Zhang, Youfang Lin, Dan Hao, and Jun Sun: **Fuzzing C++ Compilers via Type-Driven Mutation**, *OOPSLA 2025*.
16. Ruihan Zhang, and Jun Sun: **Correct-By-Construction: Certified Individual Fairness through Neural Network Training**, *OOPSLA 2025*.
17. Huan Sun, David Sanan, Jingyi Wang, Yongwang Zhao, Jun Sun, and Wenhai Wang: **Generalized Security-Preserving Refinement for Concurrent Systems**, *CCS 2025*.
18. Zhe Li, Wei Zhao, Yige Li, and Jun Sun: **"Do Influence Functions Work on Large Language Models?"**, *Findings of EMNLP 2025*.

19. Wei Zhao, Zhe Li, Yige Li, and Jun Sun: **"Zero-Shot Defense Against Toxic Images via Inherent Multimodal Alignment in LVLMs"**, *Findings of EMNLP 2025*.
20. Xinyao Xu, Ziyu Mao, Jianzhong Su, Xingwei Lin, David Basin, Jun Sun, and Jingyi Wang: **Quantitative Runtime Monitoring of Ethereum Transaction Attacks**, *WWW 2025*.
21. Songyang Yan, Xiaodong Zhang, Kunkun Hao, Haojie Xin, Yonggang Luo, Jucheng Yang, Ming Fan, Chao Yang, Jun Sun, and Zijiang Yang: **On-demand Scenario Generation for Testing Automated Driving Systems**, *FSE 2025*.
22. Yang Sun, Christopher M. Poskitt, Kun Wang, and Jun Sun: **FixDrive: Automatically Repairing Autonomous Vehicle Driving Behaviour for $0.08 per Violation**, *ICSE 2025*.
23. Ziyu Mao, Jingyi Wang, Jun Sun, Shengchao Qin, and Jiawen Xiong: **LLM-aided Automatic Modeling for Security Protocol Verification**, *ICSE 2025*.
24. Hanmo Yu, Zan Wang, Xuyang Chen, Junjie Chen, Jun Sun, Shuang Liu, and Zishuo Dong: **Mitigating Regression Faults Induced by Feature Evolution in Deep Learning Systems**, *TOSEM 2025*.
25. Junjie Chen, Xingyu Fan, Chen Yang, Shuang Liu, and Jun Sun: **De-duplicating Silent Compiler Bugs via Deep Semantic Representation**, *FSE 2025*.
26. Shunkai Zhu, Jun Sun, Jingyi Wang, Xingwei Lin, and Peng Cheng: **OptSE: Towards Optimal Symbolic Execution**, *TSE 2025*.
27. Bo Wang, Chong Chen, Junjie Chen, Bowen Xu, Chen Ye, Youfang Lin, Guoliang Dong, and Jun Sun: **A Comprehensive Study of OOP-Related Bugs in C++ Compilers**, *TSE 2025*.
28. Shunkai Zhu, Jun Sun, Jingyi Wang, Zhenbang Chen, and Peng Cheng: **Integrating Path Selection for Symbolic Execution and Variable Selection for Constraint Solving**, *TOSEM 2025*.
29. Shengping Xiao, Yongkang Li, Shufang Zhu, Jun Sun, Jianwen Li, Geguang Pu, and Moshe Vardi: **An On-the-fly Synthesis Framework for LTL over Finite Traces**, *TOSEM 2025*.
30. Nay Myat Min, Long H. Pham, and Jun Sun: **Unified Neural Backdoor Removal with Only Few Clean Samples through Unlearning and Relearning**, *TIFS 2025*.
31. Yige Li, Jiabo He, Hanxun Huang, Jun Sun, Xingjun Ma, Yu-Gang Jiang: **Shortcuts Everywhere and Nowhere: Exploring Multi-Trigger Backdoor Attacks**, *TDSC 2025*.