JIAWEN ZHANG

https://kevin-zh-cs.github.io kevinzh@zju.edu.cn

EDUCATION

Zhejiang University
PhD of Engineering in Computer Science and Technology *Sept. 2022 - Dec. 2026 (expected)*
Advised by Prof. Xiaohu Yang and Prof. Jian Liu

Zhejiang University
Bachelor of Engineering in Computer Science and Technology *Sept. 2018 - June. 2022*
GPA: 3.8/4.0

RESEARCH INTERESTS

Trustworthy LLM/Agent, AI Safety and Applied Cryptography

HIGHLIGHTED PUBLICATIONS

[1] Understanding and Preserving Safety in Fine-Tuned LLMs [paper]
Jiawen Zhang, Yangfan Hu, Kejia Chen, Lipeng He, Jiachen Ma, Jian Lou, Dan Li, Jian Liu, Xiaohu Yang, Ruoxi Jia.
- *ACM Conference on Computer and Communications Security (CCS), 2026.*

[2] Safety at One Shot: Patching Fine-Tuned LLMs with A Single Instance [paper]
Jiawen Zhang, Lipeng He, Kejia Chen, Jian Lou, Jian Liu, Xiaohu Yang, Ruoxi Jia.
- *International Conference on Learning Representations (ICLR), 2026.*

[3] Secure Transformer Inference Made Non-interactive [paper] [133 Citations]
Jiawen Zhang, Xinpeng Yang, Lipeng He, Kejia Chen, Wen-jie Lu, Yinghao Wang, Xiaoyang Hou, Jian Liu, Kui Ren, Xiaohu Yang.
- *Network and Distributed System Security (NDSS) Symposium, 2025.*
- **Normalized Top-100 Security Papers [ranking]**

[4] Activation Approximations Can Incur Safety Vulnerabilities in Aligned LLMs: Comprehensive Analysis and Defense [paper]
Jiawen Zhang*, Kejia Chen*, Lipeng He*, Jian Lou, Dan Li, Zunlei Feng, Mingli Song, Jian Liu, Kui Ren, Xiaohu Yang.
- *USENIX Security (USENIX Sec) Symposium, 2025.*

OTHER (CO-)FIRST/CORRESPONDING-AUTHOR PUBLICATIONS

[1] Beyond Similarity: Trustworthy Memory Search for Personal AI Agents [paper]
Jiawen Zhang, Kejia Chen, Jiachen Ma, Yangfan Hu, Lipeng He, Yechao Zhang, Jian Liu, Xiaohu Yang, Tianwei Zhang, Ruoxi Jia.
- *arxiv, 2026.*

[2] Reflector: Internalizing Step-wise Reflection against Indirect Jailbreaks [paper]
Jiachen Ma*, **Jiawen Zhang***, Xiangtian Li, Bo Zou, Chaochao Lu, Chao Yang.
- *International Conference on Machine Learning (ICML), 2026.*

[3] Revealing and Benchmarking the Safety Risks in Blockchain Agents
Jiawen Zhang, Kejia Chen, Lipeng He, Yechao Zhang, Jian Liu, Xiaohu Yang.

- *International Conference on Blockchain Research and Applications (BCRA), 2026.*

[4] **Assessing Safety Risks and Quantization-aware Safety Patching for Quantized Large Language Models** [paper]
Kejia Chen*, **Jiawen Zhang***, Jiacong Hu, Yu Wang, Mingli Song, Jian Lou, Zunlei Feng.
- *International Conference on Machine Learning (ICML), 2025.*

[5] **SecPE: Secure Prompt Ensembling for Private and Robust LLMs** [paper]
Jiawen Zhang*, Kejia Chen*, Zunlei Feng, Mingli Song, Jian Lou, Jian Liu, Xiaohu Yang.
- *European Conference on Artificial Intelligence (ECAI), 2024.*

[5] **SmartZKCP: Towards Practical Data Exchange Marketplace Against Active Attacks** [paper]
Xuanming Liu*, **Jiawen Zhang***, Yinghao Wang, Xinpeng Yang and Xiaohu Yang.
- *International Conference on Blockchain Research and Applications (BCRA), 2024.*

EXPERIENCE

Nanyang Technological University *Jan 2026 - July 2026*
- Visiting Scholar

Hyperchain, Hangzhou, China *Sept 2021 - June 2022*
- Blockchain Research & Design Engineer

Meituan, Shanghai, China *May 2021 - Sept 2021*
- Full Stack Develop Engineer

Ant Group, Hangzhou, China *Oct 2020 - Feb 2021*
- Full Stack Develop Engineer

SELECTED AWARDS & HONORS

China Association for Science and Technology (CAST) Youth Talent Program *2025*

China Cybersecurity Student Innovation Funding Program (Top 1%) *2025*

Outstanding Graduate Student Scholarship (Top 5%) *2025*

Outstanding Graduate Student Scholarship (Top 5%) *2024*

Zhejiang Province Outstanding Graduate Award *2022*