何嘉铭

he.jiaming@student.zy.cdut.edu.cn
17358650030
个人主页：https://leanwithming.github.io/Homepage/

教育背景

成都理工大学，软件工程，本科 2021.9 - 2025.7

- 专业排名：1/104
- 成都理工大学优秀学生，计算机与网络安全学院优秀学生

电子科技大学，网络空间安全，博士 2025.9 - 2030.7

即将入学

研究经历

电子科技大学网络安全研究院（指导老师：李洪伟，IEEE Fellow），科研助理 2023.5 - 至今

在互联网数据安全实验室工作期间，我在李洪伟教授的指导下进行人工智能安全的研究。在人工智能安全领域中，我主要关注大型基础模型的安全研究（例如，大型语言模型的越狱和基于扩散的 Text2Image 场景的后门攻击）。期间产出的工作以第一作者身份已被 CCF-A/B/C 会议 (AAAI 2025 (Oral), ICASSP 2025, ICC 2024 (Oral)) 接收。

King Abdullah University of Science and Technology (KAUST)，科研实习生（远程） 2024.5 - 至今

在 KAUST 远程科研实习期间，我主要进行关于机器遗忘（Machine Unlearning）和成员推理攻击相关（Membership Inference Attack）的研究。在主流的视觉语言多模态模型上进行相关探究（例如 CLIP 和 Llava）。

论文发表

“Watch Out for Your Guidance on Generation! Exploring Conditional Backdoor Attacks against Large Language Models” AAAI Conference on Artificial Intelligence (CCF-A) ***Oral (Top 5%)***
Jiaming He, Wenbo Jiang, Hongwei Li, Guanyu Hou, Wenshu Fan, Rui Zhang

Nov 2024 - Accepted

”Weaponizing Tokens: Backdooring Text-to-Image Generation via Token Remapping” IEEE International Conference on Multimedia & Expo (ICME) (CCF-B)
Jiaming He, Wenbo Jiang, Guanyu Hou, Qiyang Song, Hongwei Li

Mar 2025 - Accepted

”PRESS: Defending Privacy in Retrieval-Augmented Generation via Embedding Space Shifting” IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (CCF-B)
Jiaming He, Cheng Liu, Guanyu Hou, Wenbo Jiang, Jiacheng Li

Sep 2024 - Online

“MTISA: Multi-Target Image-Scaling Attack” 2024 IEEE International Conference on Communications (ICC) (CCF-C) *Oral*
Jiaming He, Hongwei Li, Wenbo Jiang, Yuan Zhang

Jan 2024 - Online

”Data Stealing Attacks against Large Language Models via Backdooring” Electronics (JCR-Q2)
Jiaming He, Guanyu Hou, Xinyue Jia, Yangyang Chen, Wenqi Liao, Yinghang Zhou, Rang Zhou

Jul 2024 - Online

”When Hallucinated Concepts Cross Modals: Unveiling Backdoor Vulnerability in Multi-modal In-context Learning” 2025 Annual Meetings of the Association for Computational Linguistics (ACL) (CCF-A)
Jiaming He, Yitong Qiao, Guanyu Hou, Wenbo Jiang, Zihan Wang, Qiyang Song, Hongwei Li

Feb 2025 - Under Review

”BadRefSR: Backdoor Attacks Against Reference-based Image Super Resolution” IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (CCF-B)
Xue Yang, Tao Chen, Lei Guo, Wenbo Jiang, Ji Guo, Yongming Li, **Jiaming He**

Sep 2024 - Online

”BadTTS: Identifying Vulnerabilities in Neural Text-to-Speech Models” IEEE Global Communications Conference (Globecom) (CCF-C)
Rui Zhang, Hongwei Li, Wenbo Jiang, Ran Zhang, **Jiaming He**

July 2024 - Accepted

”Enhancing Jailbreak Attack from the Perspective of Psychology” 2025 Annual Meetings of the Association for Computational Linguistics (ACL) (CCF-A)
Yitong Qiao, **Jiaming He**, Weibin Wu, Zhaoji Fan, Mengyu Ji, Xingxi Xian, Zibing Zheng

Feb 2025 - Under Review

”Combinational Backdoor Attacks against Customized Text-to-Image Generative models” IEEE Transactions on Forensics and Security (TIFS) (CCF-A)
Wenbo Jiang, **Jiaming He**, Hongwei Li, Rui Zhang, Wenshu Fan, Shui Yu

Jul 2024 - Under Review

“Backdoor Attacks against Image-to-Image Networks” IEEE Transactions on Dependable and Secure Computing (TDSC) (CCF-A)

Wenbo Jiang, Hongwei Li, **Jiaming He**, Rui Zhang, Guowen Xu, Tianwei Zhang, Rongxing Lu

Nov 2023 - Under Review

专业技能

编程：Python, C++, Java, Pytorch, TensorFlow

工具：LaTex, Microsoft Office

英语：英语六级