# Gelei Deng

+65-8403-0341, gelei.deng@e.ntu.edu.sg

## Profile

Ph.D. in Cybersecurity. Certified penetration tester and auditor. Skilled researcher with a robust background in computer engineering, focusing extensively on cybersecurity. Demonstrates a profound passion for and commitment to advancing security technologies and methodologies.

### Areas of Expertise

- System Security
- AI Security
- Blockchain Security
- Penetration Testing and Software Security Testing.
- Proficient in Python and Solidity; Familiar with other mainstream Languages including JavaScript, PHP, Java, C#, etc.

## Education

Nanyang Technological University (Aug 2020 – Oct 2024)

- Ph.D. Computer Science, Cybersecurity
- NTU Cyber Security Lab
- Main Research Interests on System Security, Web Security and Penetration Testing

Singapore University of Technology and Design (May 2015 – Sep 2018)

- Department of Engineering Product Design, B.E. Electrical Engineering
- Singapore Ministry of Education Full Scholarship (SM2) Holder
- Engineering Product Design Track, Honor List

## Employment

**OpenAI** (Jan 2024 – Present)

- **Position**: Independent Contractor at OpenAI Red Teaming Network.
- Participated in OpenAI led red teaming efforts to assess the risks and safety profile of OpenAI models and systems.

**Quantstamp, Inc.** (Oct 2022 – Present)

- **Position**: Lead AI Engineer, Auditing Engineer
- Conducted comprehensive security audits across multiple blockchain projects, including DeFi platforms, wallets, decentralized exchanges, and NFTs.
- Leading ongoing research on leveraging Large Language Models for automated smart contract audits and fuzzing.

**Institute for Infocomm Research ($I^2R$, A*STAR)** (Jan 2019 – Jul 2020)

- **Position**: Research Engineer, Penetration Tester.
- Perform Penetration Testing and Conduct Research Works for Singapore government-based Agencies.

## Academic Research

**Cyber Security Lab, NTU** (July 2020 - Present)

- Ph.D. student in Cybersecurity

- Main research topics include system security, robotics security, Web3 security, and penetration testing automation.
- Collaborate with industrial partners (Huawei, etc.) to verify system security and stability.

**SUTD-MIT International Design Center** (2016-2018)

- Work on hardware-oriented digital signal processing
- Leading the research work in IIR filter design
- Involved in IDCT image processing research work

**Publications**

**First-Author Publications**

**Gelei Deng,** Yi Liu, Víctor Mayoral-Vilches, Peng Liu, Yuekang Li, Yuan Xu, Tianwei Zhang, Yang Liu, Martin Pinzger, Stefan Rass, "PentestGPT: An LLM-empowered Automatic Penetration Testing Tool," in *33rd USENIX Security Symposium (USENIX '24)*. Distinguished Artifact Award, 2024. (*citation*: 63)

**Gelei Deng**, Yi Liu, Kailong Wang, Yuekang Li, Tianwei Zhang, Yang Liu, "PANDORA: Jailbreak GPTs by Retrieval Augmented Generation Poisoning," in *Workshop on Artificial Intelligence System with Confidential Computing (AISCC)*, Distinguished Paper Award, February, 2024. (*citation*: 18)

**Gelei Deng**, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, Yang Liu, "MASTERKEY: Automated Jailbreaking of Large Language Model Chatbots," in *Network and Distributed System Security Symposium (NDSS '24)*. 2024. (*citation*: 218)

**Gelei Deng**, Zhiyi Zhang, Yuekang Li, Yi Liu, Tianwei Zhang, Yang Liu, Guo Yu, Dongjin Wang, "NAUTILUS: Automated RESTful API Vulnerability Detection," in 3*2nd USENIX Security Symposium (USENIX '23)*. 2023. (*citation*: 9)

**Gelei Deng**, Guowen Xu, Yuan Zhou, Tianwei Zhang, and Yang Liu, "On the (In)Security of Secure ROS2," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22)*. Association for Computing Machinery, New York, NY, USA, 739–753. (*citation*: 19)

**Gelei Deng**, Yuan Zhou, Yuan Xu, Tianwei Zhang, and Yang Liu, "An Investigation of Byzantine Threats in Multi-Robot Systems," in *24th International Symposium on Research in Attacks, Intrusions and Defenses (RAID '21)*, New York, NY, USA, 17–32. (*citation*: 23)

**Gelei Deng**, Stefanie Yu Xingjie and Huaqun Guo, "Efficient Password Guessing based on a Password Segmentation Approach," in *GLOBECOM 2019*, Waikoloa, Hawaii, 2019. (*citation*: 14)

**Other Selected Publications**

Zihao Xu, Yi Liu, **Gelei Deng**, Yuekang Li, and Stjepan Pice, "A Comprehensive Study of Jailbreak Attack versus Defense for Large Language Models," in *Findings of the Association for Computational Linguistics ACL*, 2024. (*citation*: 4)

Yuxi Li, Yi Liu, Gelei Deng, Ying Zhang, Wenjia Song, Ling Shi, Kailong Wang, Yuekang Li, Yang Liu, and Haoyu Wang, "Glitch Tokens in Large Language Models: Categorization

Taxonomy and Effective Detection," in Proceedings of ACM Software Engineering, FSE, Article 92, July 2024. (*citation*: 8)

Xingshuo Han, Haozhao Wang, Kangqiao Zhao, **Gelei Deng**, Yuan Xu, Hangcheng Liu, Han Qiu, Tianwei Zhang, "VisionGuard: Secure and Robust Visual Perception of Autonomous Vehicles in Practice," in *ACM Conference on Computer and Communications Security (CCS)*, October, 2024. (*citation*: 0, new publication)

Yuan Xu, **Gelei Deng**, Xingshuo Han, Guanlin Li, Han Qiu, Tianwei Zhang, "PhyScout: Detecting Sensor Spoofing Attacks via Spatio-temporal Consistency," in *ACM Conference on Computer and Communications Security (CCS)*, October, 2024. (*citation*: 0, new publication)

Kunsheng Tang, Wenbo Zhou, Jie Zhang, Aishan Liu, **Gelei Deng**, Shuai Li, Peigui Qi, Weiming Zhang, Tianwei Zhang, Nenghai Yu, "GenderCARE: A Comprehensive Framework for Assessing and Reducing Gender Bias in Large Language Models," in *ACM Conference on Computer and Communications Security (CCS)*, October, 2024. (*citation*: 0, new publication)

Yi Liu*, **Gelei Deng***, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lidao Zhao, Tianwei Zhang, Yang Liu, "Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study," in *Proceedings of the 4th International Workshop on Software Engineering and AI for Data Quality in Cyber-Physical Systems/Internet of Things*. 2024. (citation: 288)

Yi Liu*, **Gelei Deng***, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, Yang Liu, "Prompt Injection Attack Against LLM-Integrated Applications," Arxiv. 2023. (citation: 180)

Ruichao Liang, Jing Chen, Kun He, Yueming Wu, **Gelei Deng**, Ruiying Du, Cong Wu, "PonziGuard: Detecting Ponzi Schemes on Ethereum with Contract Runtime Behavior Graph (CRBG)," in *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering (ICSE 2023)*, 2023. (*citation*: 8)

Yi Liu, Yuekang Li, **Gelei Deng**, Felix Juefei-Xu, Yao Du, Cen Zhang, Chengwei Liu, Yeting Li, Lei Ma, Yang Liu, "ASTER: Automatic Speech Recognition System Accessibility Testing for Stutterers," in *38th IEEE/ACM International Conference on Automated Software Engineering (ASE 2023)*. 2023. (*citation*: 3)

Yuan Xu, Xingshuo Han, **Gelei Deng**, Jiwei Li, Tianwei Zhang, Yang Liu, "SoK: Rethinking Sensor Spoofing Attacks against Robotic Vehicles from a Systematic View," in *8th IEEE European Symposium on Security and Privacy (Euro S&P 23)*. 2023. (*citation*: 20)

Yisroel Mirsky, Ambra Demontis, Jaidip Kotak, Ram Shankar, **Deng Gelei**, Liu Yang, Xiangyu Zhang, Maura Pintor, Wenke Lee, Yuval Elovici, Battista Biggio, "The Threat of Offensive AI to Organizations," *Computer & Security*, vol.124, 2023. (*citation*: 88)

Yi Liu, Yuekang Li, **Gelei Deng**, Yang Liu, et al., "Morest: Model-based RESTful API Testing with Execution Feedback," in *2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE)*, 2022. (*citation*: 32)

Yuan Xu, **Gelei Deng**, Tianwei Zhang, Han Qiu, Yungang Bao, "Novel denial-of-service attacks against cloud-based multi-robot systems," *Information Sciences*, Volume 576, 2021, Pages 329-344. (*citation*: 33)

Luying Zhou, Huaqun Guo and **Gelei Deng**, "A fog computing based approach to DDoS mitigation in IIoT systems," *Computer & Security*, vol. 85, pp. 51-62, 2019. (*citation*: 108)

**Certificates, Awards and Activities**

### Projects

PentestGPT: An LLM-empowered Automatic Penetration Testing Tool

- Open-source tool with more than 7k stars on GitHub: http://pentest-gpt.com/
- Various collaboration with industrial partners (Huawei, Bytedance, etc.)
- Active open-source community with contributors

### CVEs and Vulnerability Identification

More than verified 10 CVEs including: CVE-2021-39114, CVE-2021-30224, CVE-2021-37392, CVE-2021-37393, CVE-2021-37476, CVE-2021-37477

Multiple vulnerabilities and bugs confirmed by vendors including: Atlassian Confluence, Apache Magento, Bitbucket, SEO Panel, Spree Commerce, etc.

### Certificates

- Offensive Security Web Expert (OSWE) (2021)
- Offensive Security Certified Professional (OSCP) (2020)
- BlackHat Advanced Infrastructure Hacking Completion (2019)

### Awards

SUTD Y2015 Student Honor List (2018)

- Awarded to be one of the students with best overall performance and grade
- 12 out of more than 400

The Most Amazing SUTD Student Works of the Year (2018) – Drone ranger

- 5 selected projects out of more 100 student projects in 2018

**Additional Information**

- Fluent in English and Mandarin (spoken & written)
- Actively participating varieties of sports activities