# Xinyue Shen

*Email: xinyue.shen@cispa.de*
*Site: https://xinyueshen.me*
*Google Scholar: https://tinyurl.com/mr2kemtm*

## Research Interests

Trustworthy Artificial Intelligence; Large Language Models; Hate Speech Analysis.

## Education

2021 – 2025 **CISPA Helmholtz Center for Information Security**, Germany
Ph.D. in Computer Science

2015 – 2019 **University of Electronic Science and Technology of China (UESTC)**, China
B.Sc. in Software Engineering (Cybersecurity)

## Work Experience

2021 – 2025 **CISPA Helmholtz Center for Information Security**, Germany
Ph.D. Candidate
Advisor: Michael Backes and Yang Zhang

2019 – 2021 **Alibaba**, China
Algorithm Engineer

2019.05 – 08 **Indiana University Bloomington**, USA
Research Assistant
Advisor: Xiaojing Liao

2018.02 – 11 **Alibaba**, China
Algorithm Engineer Intern

## Selected Honors and Awards

2025 **Machine Learning and Systems Rising Star**

2025 **KAUST Rising Star in AI** (7.8%)

2024 **Heidelberg Laureate Forum Young Researcher**

2024 **Abbe Grant**, Carl-Zeiss-Stiftung Foundation

2024 **Top Reviewer**, AISec Workshop

2024 **Outstanding Popular Science Work Award**, China Science Writers Association

2022 **Chinese-Language Category Winner**, The EELISA Science Fiction Contest (3.0%)

2021 **Light-Year Award**, Beijing Association for Science and Technology (0.1%)

2019 **Valedictorian of UESTC**

2019 **Outstanding Student of UESTC**, the highest honor awarded to UESTC students (0.2%)

2017 **First Prize**, Intel National College Student Software Competition (2.0%)

2016 **First Prize**, Internet Innovation Competition of Southwest China

2015 **Excellent Volunteer**, National Games for Persons with Disabilities & National Special Olympics Games

## Publications

Note: IEEE S&P, USENIX Security, and ACM CCS are recognized as top-tier security conferences; ACL and EMNLP are top Natural Language Processing conferences; ICWSM is a prominent conference in the Social Computing domain.

[C16] **Xinyue Shen**, Yun Shen, Michael Backes, and Yang Zhang. GPTracker: A Large-Scale Measurement of Misused GPTs. In IEEE Symposium on Security and Privacy (**IEEE S&P**). IEEE, 2025. (Acceptance rate: 14.8%)

***Our findings help the platform owner take down thousands of misused GPTs***

[C15] Yicong Tan, **Xinyue Shen**, Yun Shen, Michael Backes, and Yang Zhang. On the Effectiveness of Prompt Stealing Attacks on In-The-Wild Prompts. In IEEE Symposium on Security and Privacy (**IEEE S&P**). IEEE, 2025. (Acceptance rate: 14.8%)

[C14] **Xinyue Shen**, Yixin Wu, Yiting Qu, Michael Backes, Savvas Zannettou, and Yang Zhang. HateBench: Benchmarking Hate Speech Detectors on LLM-Generated Content and Hate Campaigns. In USENIX Security Symposium (**USENIX Security**). USENIX, 2025. (Acceptance rate: TBA)

***Artifact Badges: Available, Functional, Results Reproduced***

[C13] Yihan Ma, **Xinyue Shen**, Yiting Qu, Ning Yu, Michael Backes, Savvas Zannettou, and Yang Zhang. From Meme to Threat: On the Hateful Meme Understanding and Induced Hateful Content Generation in Open-Source Vision Language Models. In USENIX Security Symposium (**USENIX Security**). USENIX, 2025. (Acceptance rate: TBA)

[C12] **Xinyue Shen**, Yun Shen, Michael Backes, Yang Zhang. When GPT Spills the Tea: Comprehensive Assessment of Knowledge File Leakage in GPTs. In Annual Meeting of the Association for Computational Linguistics (**ACL**). ACL, 2025. (Acceptance rate: TBA)

[C11] Junjie Chu, Yugeng Liu, Ziqing Yang, **Xinyue Shen**, Michael Backes, Yang Zhang. JailbreakRadar: Comprehensive Assessment of Jailbreak Attacks Against LLMs. In Annual Meeting of the Association for Computational Linguistics (**ACL**). ACL, 2025. (Acceptance rate: TBA)

[C10] Zhen Sun, Zongmin Zhang, **Xinyue Shen**, Ziyi Zhang, Yule Liu, Michael Backes, Yang Zhang, Xinlei He. Are We in the AI-Generated Text World Already? Quantifying and Monitoring AIGT on Social Media. In Annual Meeting of the Association for Computational Linguistics (**ACL**). ACL, 2025. (Acceptance rate: TBA)

[C9] **Xinyue Shen**, Yiting Qu, Michael Backes, and Yang Zhang. Prompt Stealing Attacks Against Text-to-Image Generation Models. In USENIX Security Symposium (**USENIX Security**). USENIX, 2024. (Acceptance rate: 18.3 %)

***Recognized in Microsoft Vulnerability Severity Classification for AI Systems***

***Dataset Downloaded Over 28K times***

[C8] **Xinyue Shen**, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. In ACM SIGSAC Conference on Computer and Communications Security (**CCS**). ACM, 2024. (Acceptance rate: 16.7 %)

***Top Cited Security Papers From 2024, Github Stars Over 3.1k***

[C7] Xinlei He, **Xinyue Shen**, Zeyuan Chen, Michael Backes, and Yang Zhang. MGTBench: Benchmarking Machine-Generated Text Detection. In ACM SIGSAC Conference on Computer and Communications Security (**CCS**). ACM, 2024. (Acceptance rate: 16.7 %)

***Top Cited Security Papers From 2024***

[C6] Yukun Jiang, Zheng Li, **Xinyue Shen**, Yugeng Liu, Michael Backes, and Yang Zhang. ModSCAN: Measuring Stereotypical Bias in Large Vision-Language Models from Vision and Language Modalities. In Conference on Empirical Methods in Natural Language Processing (**EMNLP**). ACL, 2024. (Acceptance rate: 37.7%)

[C5] Yihan Ma, **Xinyue Shen**, Yixin Wu, Boyang Zhang, Michael Backes, and Yang Zhang. The Death and Life of Great Prompts: Analyzing the Evolution of LLM Prompts from the Structural Perspective. In Conference on Empirical Methods in Natural Language Processing (**EMNLP**). ACL, 2024. (Acceptance rate: 37.7%)

[C4] Yukun Jiang, **Xinyue Shen**, Rui Wen, Zeyang Sha, Junjie Chu, Yugeng Liu, Michael Backes, and Yang Zhang. Games and Beyond: Analyzing the Bullet Chats of Esports Livestreaming. In International Conference on Web and Social Media (**ICWSM**). AAAI, 2024. (Acceptance rate: 20.0%)

[C3] Yiting Qu, **Xinyue Shen**, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafe Diffusion: On the Generation of Unsafe Images and Hateful Memes From Text-To-Image Models. In ACM SIGSAC Conference on Computer and Communications Security (**CCS**). ACM, 2023. (Acceptance rate: 19%)

[C2] **Xinyue Shen**, Xinlei He, Michael Backes, Jeremy Blackburn, Savvas Zannettou, and Yang Zhang. On Xing Tian and the Perseverance of Anti-China Sentiment Online. In International Conference on Web and Social Media (**ICWSM**). AAAI, 2022. (Acceptance rate: 22.0%)

**Before Ph.D.**

[C1] Liya Su*, **Xinyue Shen*** (* co-first author), Xiangyu Du, Xiaojing Liao, XiaoFeng Wang, Luyi Xing, and Baoxu Liu. Evil Under the Sun: Understanding and Discovering Attacks on Ethereum Decentralized Applications. In USENIX Security Symposium (**USENIX Security**). USENIX, 2021. (Acceptance rate: 18.7%)

## Industry Impacts & Media Coverage

Feb 27, 2025 OpenAI. *OpenAI GPT-4.5 System Card.*

Jan 31, 2025 OpenAI. *OpenAI o3-mini System Card.*

Jan 21, 2025 German Federal Office for Information Security (BSI). *Generative AI Models: Opportunities and Risks for Industry and Authorities.*

Oct 28, 2024 CISPA News. *Prompt Stealing: CISPA Researcher Discovers New Attack Scenario for Text-To-Image Generation Models.*

Oct 16, 2024 Spektrum.de. *At the 11th Heidelberg Laureate Forum, Young Researchers Step Into the Spotlight.*

Sep 12, 2024 OpenAI. *OpenAI o1 System Card.*

Jun 01, 2024 The Decoder. *Creative Stories Can Jailbreak ChatGPT Voice, Study Finds.*

May 30, 2024 TheCyberExpress. *Japanese Man Arrested for GenAI Ransomware as AI Jailbreak Concerns Grow.*

Jan 02, 2024 NIST. *NIST Trustworthy and Responsible AI Taxonomy.*

Aug 23, 2023 Deutschlandfunk Nova. *Wie Chatbots Die Eigenen Regeln Vergessen (How Chatbots Forget Their Own Rules.*)

Aug 21, 2023 New Scientist. *Tricks for Making AI Chatbots Break Rules Are Freely Available Online.*

Jul 26, 2023 Montreal AI Ethics Institute. *On the Generation of Unsafe Images and Hateful Memes From Text-To-Image Models.*

## Invited Talks

2025 Jul **Max Planck Institute for Security and Privacy (MPI-SP)**
Understanding and Mitigating LLM Misuse in the Real World

2025 Jul **Wuhan University**
When LLMs Are in the Wrong Hands

2025 Jul **CNIL Privacy Research Day**
HateBench: Benchmarking Hate Speech Detectors on LLM-Generated Content and Hate Campaigns

2025 Jun **LLMApp Workshop @FSE 2025**
GPTracker: A Large-Scale Measurement of Misuse and Knowledge File Leakage in GPTs

2025 Jun **Leiden University**
When LLMs Are in the Wrong Hands

2025 Jun **Delft University of Technology (TU Delft)**
When LLMs Are in the Wrong Hands

2025 May **MLCommons ML and Systems Rising Stars Workshop**
"Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models

2025 Apr **King Abdullah University of Science and Technology (KAUST)**
Understand and Mitigate AI System Misuse in the Real World

2024 Oct **Heidelberg Laureate Forum (HLF)**
"Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models

2024 Sep **The Ohio State University**
Emerging Attacks in the Era of Generative AI

2024 Jun **AEGIS Symposium on Cyber Security**
Emerging Attacks in the Era of Generative AI

2024 Apr **Google**
Emerging Attacks in the Era of Generative AI

2023 Oct **Shanghai Jiao Tong University**
Understanding and Quantifying the Safety Issues of Large Foundation Models

2023 Oct **Fudan University**
Understanding and Quantifying the Safety Issues of Large Foundation Models

2023 Sep **Sichuan University**
Understanding and Quantifying the Safety Issues of Large Foundation Models

2023 Sep **University of Electronic Science and Technology of China**
Understanding and Quantifying the Safety Issues of Large Foundation Models

2023 Jun **AEGIS Symposium on Cyber Security**
Measuring the Reliability of ChatGPT

2018 Oct **Hack In The Box Conference (HITBConf)**
Solving The Last Mile Problem Between Machine Learning and Security Operations

## Teaching & Mentoring

### Teaching

Guest Lecturer, Data-driven Understanding of the Disinformation Epidemic, CISPA 2025
Guest Lecturer, Attacks Against Machine Learning Models, CISPA 2025
Guest Lecturer, Machine Learning Security & Privacy, HKUST (Guangzhou) 2024
Guest Lecturer, Privacy of Machine Learning, CISPA 2024
Guest Lecturer, Attacks Against Machine Learning Models, CISPA 2024
Guest Lecturer, Data-driven Understanding of the Disinformation Epidemic, CISPA 2024
Teaching Assistant, Privacy of Machine Learning, CISPA 2023
Teaching Assistant, Attacks Against Machine Learning Models, CISPA 2023
Teaching Assistant, Data-driven Understanding of the Disinformation Epidemic, CISPA 2023
Teaching Assistant, Privacy of Machine Learning, CISPA 2022

### Mentoring

| | |
|---|---|
| Yicong Tan, Ph.D. Computer Science, CISPA | 2024–Present |
| Dora Chen, Ph.D. Computer Science, CISPA | 2024–Present |
| Yukun Jiang, Ph.D. Computer Science, CISPA | 2023–Present |
| Yage Zhang, M.S. Computer Science, Saarland University | 2025 |
| Thomas Boisvert, B.S. Computer Science, Saarland University | 2023 |

## Academic Service

### Conference Reviewing

| | |
|---|---|
| PC, USENIX Security Symposium (USENIX) | 2025 |
| PC, Association for Computational Linguistics (ACL) | 2025 |
| PC, International AAAI Conference on Web and Social Media (ICWSM) | 2024, 2025, 2026 |
| PC, IEEE Secure and Trustworthy Machine Learning (SaTML) | 2025, 2026 |
| PC, ACM Workshop on Artificial Intelligence and Security (AISec) | 2024, 2025 |
| Poster PC, IEEE Symposium on Security and Privacy (S&P) | 2023, 2024, 2025 |
| Poster PC, USENIX Security Symposium (USENIX) | 2024 |
| AEC, ACM Conference on Computer and Communications Security (CCS) | 2024 |
| Reviewer, ACM Conference on Human Factors in Computing Systems (CHI) | 2024 |

### Journal Reviewing

| | |
|---|---|
| Reviewer, Nature Human Behaviour | 2025 |
| Reviewer, Transactions on Information Forensics & Security (TIFS) | 2025 |
| Reviewer, ACM Transactions on Privacy and Security (TOPS) | 2024, 2025 |
| Reviewer, Transactions on Software Engineering (TSE) | 2024 |
| Reviewer, Information Processing & Management (IP&M) | 2024 |

### Organizer

| | |
|---|---|
| Organizer, LAMPS workshop at ACM CCS | 2024 |