Xinyue Shen

Email: *xinyue.shen@cispa.de*
Site: *https://xinyueshen.me*
Last update: October 8, 2025

Research Interests

Trustworthy ML, LLM/VLM Security and Safety, Social Computing, Agentic AI

Education

Apr 2021 – Present **CISPA Helmholtz Center for Information Security**, Germany
Ph.D. Candidate in Computer Science
Thesis: Understanding and Mitigating the Misuse of Real-World AI Systems
Advisors: Michael Backes and Yang Zhang

Sep 2015 – Jun 2019 **University of Electronic Science and Technology of China**, China
Bachelor in Software Engineering (Cybersecurity) with honors

Work Experience

Apr 2021 – Present **CISPA Helmholtz Center for Information Security**, Germany
Ph.D. Candidate

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

May 2019 – Aug 2019 **Indiana University Bloomington**, USA
Research Assistant
Advisor: Xiaojing Liao

Feb 2018 – Nov 2018 **Alibaba**, China
Algorithm Engineer Intern

Selected Honors and Awards

2025 **Best Machine Learning and Security Paper in Cybersecurity Award**
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.

Google Scholar: https://scholar.google.com/citations?user=N4y3p8kAAAAJ

H-index: 11, ORCID: 0009-0006-9954-587X

[C17] **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

[C16] 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%)

[C15] **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: 16.2%)

Artifact Badges: Available, Functional, Results Reproduced

[C14] 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: 16.2%)

[C13] **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: 20.3%)

[C12] 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 Oral**). ACL, 2025. (Acceptance rate: 20.3%)

[C11] 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: 20.3%)

[C10] Yiting Qu, **Xinyue Shen**, Yixin Wu, Michael Backes, Savvas Zannettou, Yang Zhang. UnsafeBench: Benchmarking Image Safety Classifiers on Real-World and AI-Generated Images. In ACM SIGSAC Conference on Computer and Communications Security (**CCS**). ACM, 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%)

Best Machine Learning and Security Paper in Cybersecurity Award 2025, 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: 20.8%)

[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: 20.8%)

[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: 19.0%)

[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

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

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

Jul 2025 **Commission Nationale de l'Informatique et des Libertés (CNIL, France's Data Protection Authority)**
HateBench: Benchmarking Hate Speech Detectors on LLM-Generated Content and Hate Campaigns

Jun 2025 **LLMApp Workshop @FSE 2025**

GPTracker: A Large-Scale Measurement of Misuse and Knowledge File Leakage in GPTs

Jun 2025 **Leiden University**

When LLMs Are in the Wrong Hands

Jun 2025 **Delft University of Technology (TU Delft)**

When LLMs Are in the Wrong Hands

May 2025 **MLCommons ML and Systems Rising Stars Workshop**

"Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models

Apr 2025 **King Abdullah University of Science and Technology (KAUST)**

Understand and Mitigate AI System Misuse in the Real World

Oct 2024 **Heidelberg Laureate Forum (HLF)**

"Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models

Sep 2024 **The Ohio State University**

Emerging Attacks in the Era of Generative AI

Jun 2024 **AEGIS Symposium on Cyber Security**

Emerging Attacks in the Era of Generative AI

Apr 2024 **Google**

Emerging Attacks in the Era of Generative AI

Oct 2023 **Shanghai Jiao Tong University**

Understanding and Quantifying the Safety Issues of Large Foundation Models

Oct 2023 **Fudan University**

Understanding and Quantifying the Safety Issues of Large Foundation Models

Sep 2023 **Sichuan University**

Understanding and Quantifying the Safety Issues of Large Foundation Models

Sep 2023 **University of Electronic Science and Technology of China**

Understanding and Quantifying the Safety Issues of Large Foundation Models

Jun 2023 **AEGIS Symposium on Cyber Security**

Measuring the Reliability of ChatGPT

Oct 2018 **Hack In The Box Conference (HITBConf)**

Solving The Last Mile Problem Between Machine Learning and Security Operations

Teaching & Mentoring

Teaching

Summer 2025 Co-Lecturer, Data-driven Understanding of the Disinformation Epidemic, CISPA

Summer 2025 Co-Lecturer, Attacks Against Machine Learning Models, CISPA

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

Mentoring

Yiru Wang, Ph.D. student, Computer Science, CUHK	2025–Present
Zhen Sun, Ph.D. student, Computer Science, HKUST (Guangzhou)	2025–Present
Xin Wang, Ph.D. student, Computer Science, Xi'an Jiaotong University	2025–Present
Ziqing Yang, Ph.D. student, Computer Science, CISPA	2025–Present
Dora Chen, Ph.D. student, Computer Science, CISPA	2024–Present
Yicong Tan, Ph.D. student, Computer Science, CISPA	2024–2025
Yukun Jiang, Ph.D. student, Computer Science, CISPA	2023–2025
Ye Leng, Ph.D. student, Computer Science, CISPA	2024–2025
Yage Zhang, M.S., Saarland University → Ph.D. student, CISPA	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, 2026

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

Organizing and Chairing

Session Chair, USENIX Security Symposium (USENIX)	2025
Organizer, LAMPS workshop at ACM CCS	2024

Science Outreach and Other Publications

With a long-term commitment to science communication, I have dedicated myself to making cutting-edge knowledge in Computer Science, AI, and Cybersecurity accessible to the general public. Since 2016, I have published 49 research-based popular science articles and science fiction works.

2025 Autonomous Driving: From Science Fiction to Reality. *Explore*, 2025(9).

2025 Help! I've Fallen into the Computer! Publishing House of Electronics Industry.

2025 Steam, Gears, and Computers: the Old Dream of Mechanical Computing. *Science Fiction World Pictorial·Amazing Science*, 2025(7–8).

2025 How Machines "See" the World. *Explore*, 2025(3).

2024 Two Years Since ChatGPT Was Released, Has the World Changed? *Explore*, 2024(9).

2024 How Far Are We from "Westworld"? *Science Fiction World Pictorial·Amazing Science*, 2024(3).

2024 Back Home. *Exploration Discovery*, 2024(1–2).

2023 The Appearance of the Hacker. *Exploration Discovery*, 2023(12).

2023 Sorrows from the Site Administrator. *Exploration Discovery*, 2023(11).

2023 Disappeared Webpage. *Exploration Discovery*, 2023(10).

2023 The Firewall and Hacker Attack. *Exploration Discovery*, 2023(9).

2023 AI Content Detective: Who is Writing the Essay? *Young Adult Computer World*, 2023(7–8).

2023 The Dream of Web Crawler. *Exploration Discovery*, 2023(8).

2023 Internet Hijacking Incident. *Exploration Discovery*, 2023(7).

2023 Strange Destination. *Exploration Discovery*, 2023(6).

2023 Go to the Internet. *Exploration Discovery*, 2023(5).

2023 Foreseeing Sadness and Heartfelt Desire: The Ways and Costs of Cross-Species Communication. *Science Fiction Cube*, 2023(5).

2023 OMG! Grandpa Domain Name System (DNS) turned out to be! *Exploration Discovery*, 2023(4).

2023 The Beginning of the Internet. *Exploration Discovery*, 2023(3).

2023 Voice Message from Mom. *Exploration Discovery*, 2023(1–2).

2023 Hacking Storm. *Exploration Discovery*, 2023(1–2).

2022 Real and Fake Process. *Exploration Discovery*, 2022(12).

2022 The Adventures on Hard Disk Island. *Exploration Discovery*, 2022(11).

2022 The Data Cable Is the Tunnel and Bluetooth Is the Ferry. *Exploration Discovery*, 2022(10).

2022 Is Artificial Intelligence a "Tower"? *Science Fiction World (Youth)*, 2022(9).

2022 Where Is the Hard Drive? *Exploration Discovery*, 2022(9).

2022 How We Communicate With Life on Earth. *CDSTM*, 2022(8).

2022 CPU: A Math Genius With Goldfish-Style Memory. *Exploration Discovery*, 2022(8).

2022 Encounter With the Process. *Exploration Discovery*, 2022(7).

2022 The Secrets Behind Sports Games. *Science Fiction World (Youth)*, 2022(7).

2022 Wake Up, Apps! *Exploration Discovery*, 2022(6).

2022 Help! I Fell Into the Computer. *Exploration Discovery*, 2022(5).

2022 Newcomers, Careful. *Science Fiction World (Youth)*, 2022(5).

2022 "Magic" Electronics. *Science Fiction World (Youth)*, 2022(3).

2022 Empty Yellow Crane Tower Here. *The EELISA Science Fiction Contest, Chinese-Language Category Winner.*

2022 Omnipotent "Little Ear". *Science Fiction World (Youth)*, 2022(1).

2022 When Trojan Virus Meets Military Training. *Exploration Discovery*, 2022(1–2). *Outstanding Popular Science Work Award, China Science Writers Association, 2024.*

2021 Anti-Flood Action on the Internet. *Science Fiction World (Youth)*, 2021(10).

2021 A Hijacked DNS Tree. *Science Fiction World (Youth)*, 2021(9).

2021 Be a Postman in Cyberspace. *Science Fiction World (Youth)*, 2021(8).

2021 A War in Computer Castle. *Science Fiction World (Youth)*, 2021(7).

2021 Double Hacker: The Darling of Cyberpunk. *Science Fiction Cube*, 2021(5).

2021 Lady White Bone. *The Ninth "Light-Year" Award, First Prize*.

2020 Hack! A Seven-Day Invasion Diary of Trojan Horse. *"Pop-science and Sci-fiction Youth Star" Award, China Science Writers Association*.

2020 Stars on the Wrist. *Science Fiction Cube*, 2020(7).

2019 A War without Smoke: the Evolution History of Hacker Empire. *Science Fiction World*, 2019(6).

2018 A Man I Am, an Apology I Owe. *Science Fiction Cube*, 2018(12).

2017 Giant Mythical Bird. *Jiuzhou Future*, 2017(1).

2016 Nice to Meet You. *Philosophy*, 2016(1).