

Yiting Qu — CV

✉ yiting.qu@cispa.de • 🌐 yitingqu.github.io

last update: August 30, 2026

Employment

CISPA Helmholtz Center for Information Security

Saarbrücken, Germany

Postdoctoral Researcher

11/2025 -

Hosted by Prof. Michael Backes and Prof. Yang Zhang

Education

CISPA Helmholtz Center for Information Security

Saarbrücken, Germany

Ph.D. in Computer Science (Awarded by Saarland University)

11/2021 - 10/2025

Advisors: Prof. Michael Backes, Prof. Yang Zhang

Thesis: Mitigating Risks in Real-World and AI-Generated Visual Content

Shanghai Jiao Tong University

Shanghai, China

M.Sc. in Economics and Management

9/2018 - 6/2021

Advisor: Prof. Suguo Du

Shandong University

Jinan, China

B.Sc. in Management

9/2014 - 6/2018

Advisor: Prof. Tao Sun

Research Interests

- **Real-world misuse of generative AI:** unsafe image generation, harmful multimodal content, hateful memes, and AI-generated hate speech.
- **Security of generative models:** prompt stealing, reasoning extraction, jailbreaking, adversarial bypasses, and cross-modal safety gaps in vision-language models.
- **Mitigation and synthetic media forensics:** safety benchmarks, harmful-content detection, deepfake analysis, and fake point-cloud detection.

Publications

My full publication list is available on DBLP and Google Scholar (600+ citations as of July 2026). My ORCID is 0009-0000-1638-7287. IEEE S&P, CCS, USENIX Security, and NDSS are widely recognized as top-tier venues in information security.

Conference.....

- [1] Bo Shao, Yihan Ma, **Yiting Qu**, Michael Backes, and Yang Zhang. MemeTrace: A Multimodal LLM-Based Agent for Safety-Aware Meme Provenance Analysis. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) Findings*. ACL, 2026 (**CCF-B, Acceptance rate: 29.7%**).

- [2] **Yiting Qu**, Xinyue Shen, Yixin Wu, Michael Backes, Savvas Zannettou, and 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 (CCF-A, Acceptance rate: 14.5%).
Integrated into Promptfoo, an AI safety testing platform under OpenAI.
- [3] **Yiting Qu**, Ziqing Yang, Yihan Ma, Michael Backes, and Yang Zhang. Hate in Plain Sight: On the Risks of Moderating AI-Generated Hateful Illusions. In *IEEE International Conference on Computer Vision (ICCV)*. ICCV, 2025 (CCF-A, Acceptance rate: 24.0%).
- [4] **Yiting Qu**, Michael Backes, and Yang Zhang. Bridging the Gap in Vision Language Models in Identifying Unsafe Concepts Across Modalities. In *USENIX Security Symposium (USENIX Security)*. USENIX, 2025 (CCF-A, Acceptance rate: 17.1%).
- [5] 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 (CCF-A, Acceptance rate: 17.1%).
Recognized by OpenAI's Promptfoo as a documented LLM vulnerability.
- [6] 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 (CCF-A, Acceptance rate: 17.1%).
- [7] **Yiting Qu**, Zhikun Zhang, Yun Shen, Michael Backes, and Yang Zhang. FAKEPCD: Fake Point Cloud Detection via Source Attribution. In *ACM Asia Conference on Computer and Communications Security (ASIACCS)*. ACM, 2024 (CORE A, Acceptance rate: 22.0%).
- [8] 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 (CCF-A, Acceptance rate: 19.2%).
Recognized by Microsoft in the Microsoft Vulnerability Severity Classification; over 30K dataset downloads.
- [9] **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 (CCF-A, Acceptance rate: 19.2%).
Cited in policy documents from the US, the UK, and Germany; Covered by Informationsdienst Wissenschaft (idw), The Next Web, and Montreal AI Ethics Institute.
- [10] **Yiting Qu**, Xinlei He, Shannon Pierson, Michael Backes, Yang Zhang, and Savvas Zannettou. On the Evolution of (Hateful) Memes by Means of Multimodal Contrastive Learning. In *IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2023 (CCF-A, Acceptance rate: 17.4%).

Journal.....

- [11] **Yiting Qu**, Suguo Du, Shaofeng Li, Yan Meng, Le Zhang, and Haojin Zhu. Automatic Permission Optimization Framework for Privacy Enhancement of Mobile Applications. *IEEE Internet of Things Journal*, 2020 (JCR Q1, IF=9.6).

Technical Report.....

- [12] **Yiting Qu**, Ziqing Yang, Chi Cui, Ye Leng, Junjie Chu, and Yang Zhang. EchoCoT: Extracting Hidden Chain-of-Thought from Large Reasoning Models. *CoRR abs/2608.20055*, 2026.
- [13] Ye Leng, Junjie Chu, **Yiting Qu**, Mingjie Li, Yun Shen, and Yang Zhang. Innocent Panels, Hateful Stories: Evaluating and Detecting Hateful Intent in Multi-Turn Visual Story Generation. *CoRR abs/2608.05210*, 2026.
- [14] Junjie Chu, **Yiting Qu**, Ye Leng, Michael Backes, Yun Shen, Savvas Zannettou, and Yang Zhang. Understanding LLM Behavior When Encountering User-Supplied Harmful Content in Harmless Tasks. *CoRR abs/2603.11914*, 2026.
- [15] Mengfei Liang, **Yiting Qu**, Yukun Jiang, Michael Backes, and Yang Zhang. From Evidence to Verdict: An Agent-Based Forensic Framework for AI-Generated Image Detection. *CoRR abs/2511.00181*, 2025.

Research Impact and Media Coverage

- 2026 May - UnsafeBench was integrated into *Promptfoo* under OpenAI as an official plugin.
- 2025 Oct - UK Department for Science, Innovation & Technology (DSIT). *International Scientific Report on the Safety of Advanced AI: Interim Report*
- 2025 Jan - German Federal Office for Information Security (BSI). *Generative AI Models: Opportunities and Risks for Industry and Authorities*.
- 2025 Jan - HateBench was incorporated into *Promptfoo* as an LLM hate campaign vulnerability.
- 2024 July - US. National Institute of Standards and Technology (NIST). *NIST Trustworthy and Responsible AI: NIST AI 600-1*
- 2024 July - Unsafe Diffusion was included in *Awesome-Diffusion-Models* with 11K+ GitHub stars.
- 2024 July - NetApp. *FAKEPCD: Fake Point Cloud Detection via Source Attribution*.
- 2024 Jan - Prompt stealing attack was recognized in the *Microsoft Vulnerability Severity Classification for AI Systems*.
- 2024 Jan - US. National Institute of Standards and Technology (NIST). *NIST Trustworthy and Responsible AI: NIST AI 100-2e2023*
- 2023 Oct - Informationsdienst Wissenschaft (idw). *AI Image Generators as Drivers of Unsafe Images? CIPA Researcher Develops Filter to Tackle This*.
- 2023 Jul - Montreal AI Ethics Institute (MAIEI). *On the Generation of Unsafe Images and Hateful Memes From Text-To-Image Models*.

Selected Honors and Awards

- 2024 - Excellence Scholarship for Overseas Self-financed PhD Students
- 2018 - National Scholarship at Shanghai Jiao Tong University
- 2016 - Weichai Scholarship at Shandong University

Teaching

- 2025, Co-Lecturer, Attacks Against Machine Learning Models, CISPA
- 2025, Teaching Assistant, Data-driven Understanding of the Disinformation Epidemic, CISPA
- 2024, Co-Lecturer, Attacks Against Machine Learning Models, CISPA
- 2024, Teaching Assistant, Data-driven Understanding of the Disinformation Epidemic, CISPA
- 2020, Teaching Assistant, Statistics and Machine Learning, Shanghai Jiao Tong University

Mentoring

- 2026-2027, Ye Leng, Ph.D. student, Computer Science, CISPA
- 2025-2026, Bo Shao, Ph.D. student, Computer Science, CISPA
- 2025-2026, Mengfei Liang, Ph.D. student, Computer Science, CISPA
- 2025-2026, Yibo Liang, Ph.D. student, Computer Science, Xi'an Jiaotong University
- 2025-2026, Yuheng Wang, B.S., Computer Science, Xi'an Jiaotong University
- 2024-2025 Junjie Chu, Ph.D. student, Computer Science, CISPA

Academic Service

Program Committees.....

- USENIX Security Symposium, 2027
- The Association for the Advancement of Artificial Intelligence (AAAI), 2027
- Privacy Enhancing Technologies Symposium (PETs/PoPETs), 2026, 2027
- Conference on Machine Learning and Systems (MLSys), 2026
- USENIX Security Symposium, Artifact Evaluation Committee (USENIX Security AEC), 2025

Journal Reviewing.....

- IEEE Transactions on Information Forensics and Security (TIFS), 2025, 2026
- IEEE Transactions on Dependable and Secure Computing (TDSC), 2025
- ACM Transactions on Privacy and Security (TOPS), 2024

Conference Reviewing.....

- 2026: KDD, ICML, USENIX Security
- 2025: CCS, USENIX Security, ACL, ICCV, WWW
- 2024: IEEE S&P, CCS, USENIX Security, CVPR, WWW, KDD, ECCV, ICLR