Yiting Qu — CV

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

Employment

CISPA Helmholtz Center for Information Security — **Saarbrücken, Germany**
Postdoc — *7/2025 -*
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 - 6/2025*
Advisors: Prof. Michael Backes, Prof. Yang Zhang

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

- Trustworthy Machine Learning (Privacy, Security, and Safety)
- Safety of Large Foundation Models
- Online Hate, Memes, and Misinformation
- Social Network Analysis

Publications

My publication list can also be found at DBLP and Google Scholar. Note that in the domain of information security, the most prestigious conferences are IEEE S&P, CCS, USENIX Security, and NDSS.

Conference

[1] **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.

[2] **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.

[3] **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.

[4] 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.

[5] 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.

[6] **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.

[7] 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.

[8] **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.

[9] **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.

Journal

[10] **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.

Media Coverage

- 2025 Jan - German Federal Office for Information Security (BSI). *Generative AI Models: Opportunities and Risks for Industry and Authorities.*
- 2024 Oct - CISPA News. *Prompt Stealing: CISPA Researcher Discovers New Attack Scenario for Text-to-Image Generation Models.*
- 2024 Oct - Medium. *Towards Safer Visual Language Models: A Review of Current Safety Evaluation Benchmarks.*
- 2024 July - NetApp. *FAKEPCD: Fake Point Cloud Detection via Source Attribution.*
- 2023 Oct - Informationsdienst Wissenschaft (idw). *AI Image Generators as Drivers of Unsafe Images? CISPA 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.*

Invited Talks

- 2024 Jan - **Fudan University**
  The Dual Roles of Multimodal Models

- 2024 Jan - **Shanghai Jiao Tong University**
  The Dual Roles of Multimodal 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

Service

- Program Committee Member
  - 2025: USENIX Security (AEC), KDD (Research Track)
- Conference Reviewer
  - 2025: WWW, ACL, ICCV, CCS, USENIX
  - 2024: IEEE S&P, CCS, ECCV, CVPR, WWW, KDD, ICLR
  - 2023: CCS, NDSS, NeurIPS, WWW, KDD, AISec, SaTM
  - 2022: CCS, AsiaCCS, AISec
- Journal Reviewer
  - 2025: TDSC, TIFS
  - 2024: TOPS

Teaching Assistant

- Advanced Lecture: Attacks Against Machine Learning Models (2025 Summer)
- Seminar: Data-driven Understanding of the Disinformation Epidemic (2025 Summer)
- Advanced Lecture: Attacks Against Machine Learning Models (2024 Summer)
- Seminar: Data-driven Understanding of the Disinformation Epidemic (2024 Summer)
- Advanced Lecture: Statistics and Machine Learning (2020 Winter)