Yiting Qu — CV

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

Education

CISPA Helmholtz Center for Information Security — **Saarbrücken, Germany**
Ph.D. in Computer Science (Awarded by Saarland University) — *11/2021 - 6/2025 (Estimated)*
Advisors: Prof. Michael Backes, Prof. Yang Zhang

Shanghai Jiao Tong University — **Shanghai, China**
Master in Economics and Management — *9/2018 - 6/2021*
Advisor: Prof. Suguo Du

Shandong University — **Jinan, China**
Bachelor in Management — *9/2014 - 6/2018*
Advisor: Prof. Tao Sun

Research Interests

- Trustworthy Machine Learning (Privacy, Security, and Safety)
- Online Hate, Memes, and Misinformation
- Social Network Analysis
- AI for Science

Publication

Conference

[1] 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.

[2] 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.

[3] **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.

[4] 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.

[5] **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.

[6] **Yiting Qu** and Xinlei He and Shannon Pierson and Michael Backes and 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

[7] **Yiting Qu** and Suguo Du and Shaofeng Li and Yan Meng and Le Zhang and Haojin Zhu. Automatic Permission Optimization Framework for Privacy Enhancement of Mobile Applications. *IEEE Internet of Things Journal*, 2020.

Technical Report

[8] **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. *CoRR abs/2405.03486*, 2024.

Service

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

Teaching Assistant

- 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)