

# YUGENG LIU

Stuhlsatzenhausweg 5, 66123, Saarbrücken, Germany

yugeng.liu@cispa.de  $\diamond$  <https://liu.ai>

## EDUCATION

---

**CISPA Helmholtz Center for Information Security**

Since 01/2022

*Ph.D. Student*

*Supervisor: Michael Backes and Yang Zhang*

**Saarland University**

10/2019 - 07/2021

*Preparatory Phase*

**Shanghai Jiao Tong University**

09/2014 - 07/2018

*Bachelor in Computer Science and Technology*

## PUBLICATIONS

---

**Robustness Over Time: Understanding Adversarial Examples' Effectiveness on Longitudinal Versions of Large Language Models**

Yugeng Liu, Tianshuo Cong, Zhengyu Zhao, Michael Backes, Yun Shen, Yang Zhang. *TIFS*, 2026.

**Amplifying Machine Learning Attacks Through Strategic Compositions**

Yugeng Liu, Zheng Li, Hai Huang, Michael Backes, Yang Zhang. *Preprint*, 2025.

**Watermarking LLM-Generated Datasets in Downstream Tasks**

Yugeng Liu, Tianshuo Cong, Michael Backes, Zheng Li, Yang Zhang. *Preprint*, 2025.

**JailbreakRadar: Comprehensive Assessment of Jailbreak Attacks Against LLMs**

Junjie Chu, Yugeng Liu, Ziqing Yang, Xinyue Shen, Michael Backes, Yang Zhang. *ACL*, 2025.

**Neeko: Model Hijacking Attacks Against Generative Adversarial Networks**

Junjie Chu, Yugeng Liu, Xinlei He, Michael Backes, Yang Zhang, Ahmed Salem. *ICME*, 2025.

**ModSCAN: Measuring Stereotypical Bias in Large Vision-Language Models from Vision and Language Modalities**

Yukun Jiang, Zheng Li, Xinyue Shen, Yugeng Liu, Michael Backes, Yang Zhang. *EMNLP*, 2024.

**Games and Beyond: Analyzing the Bullet Chats of Esports Livestreaming**

Yukun Jiang, Xinyue Shen, Rui Wen, Zeyang Sha, Junjie Chu, Yugeng Liu, Michael Backes, Yang Zhang. *ICWSM*, 2024.

**Watermarking Diffusion Model**

Yugeng Liu, Zheng Li, Michael Backes, Yun Shen, Yang Zhang. *Preprint*, 2023.

**Backdoor Attacks Against Dataset Distillation**

Yugeng Liu, Zheng Li, Michael Backes, Yun Shen, Yang Zhang. *NDSS*, 2023.

**ML-Doctor: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models**

Yugeng Liu\*, Rui Wen\*, Xinlei He, Ahmed Salem, Zhikun Zhang, Michael Backes, Emiliano De Cristofaro, Mario Fritz, Yang Zhang. *USENIX Security*, 2022 (\* equal contribution).

**Securing Android Applications via Edge Assistant Third-Party Library Detection**

Zhushou Tang, Minhui Xue, Guozhu Meng, Chengguo Ying, Yugeng Liu, Jianan He, Haojin Zhu, Yang Liu. *Computers & Security*, 2019.

## HoMonit: Monitoring SmartHome Apps from Encrypted Traffic

Wei Zhang, Yan Meng, **Yugeng Liu**, Xiaokuan Zhang, Yinqian Zhang, Haojin Zhu. *CCS*, 2018.

## SERVICES

---

Conference PC Member:

- 2026: IEEE SaTML
- 2025: IEEE Euro S&P, IEEE SaTML

Conference Reviewer:

- 2026: ACM KDD, IEEE ICME, ICML
- 2025: AAAI, ACM KDD, IEEE ICME
- 2024: ACM WWW
- 2022: SocInfo

## RESEARCH INTERNSHIP

---

Nokia Bell Lab	07/2024 - 10/2024
The Johns Hopkins University	06/2019 - 09/2019
Shanghai Benzhong Information Technology Co., Ltd.	09/2017 - 01/2018

## HONORS & AWARDS

---

TDSC Reviewer Certificate 2025

## INVITED TALKS

---

2023.09: Secure Machine Learning, Xidian University, China.

2024.06: Trustworthy Large Language Models, Zhejiang University, China.