

Yiyong Liu

PH.D. CANDIDATE IN COMPUTER SCIENCE · MACHINE LEARNING SECURITY & PRIVACY

Saarbrücken, Germany

✉ yiyong.liu@cispa.de | 🏠 liu199604.github.io | 📷 Liu199604 | 🎓 Yiyong Liu

Summary

Final-year Ph.D. student working on the security and privacy of machine learning, with first-author publications at CCS 2022 and USENIX Security 2025. Research spans privacy attacks on machine learning models (membership inference, data reconstruction) and the safety of large language models, including jailbreak defenses for reasoning models and the security of AI agents.

Research Interests

- Trustworthy Machine Learning (Security, Privacy, and Robustness)
- Security and Privacy of Large Reasoning Models
- Safety and Alignment of AI Agents
- Adversarial Machine Learning

Education

CISPA Helmholtz Center for Information Security & Saarland University

PH.D. IN COMPUTER SCIENCE (EXPECTED MARCH 2027)

- Advisor: Prof. Dr. Michael Backes

Saarbrücken, Germany

Oct. 2022 - Present

CISPA Helmholtz Center for Information Security & Saarland University

PH.D. PREPARATORY PHASE IN COMPUTER SCIENCE

Saarbrücken, Germany

Apr. 2021 - Sep. 2022

Zhejiang University

B.ENG. IN CONSTRUCTION ENGINEERING

Hangzhou, China

Aug. 2015 - Jun. 2019

Academic Experience

Institute of Science Tokyo

VISITING RESEARCHER

- Collaborated with Prof. Jun Sakuma on AI safety, proposing an inference-time guardrail framework (CRG) to defend Large Reasoning Models against reasoning-centric jailbreaks via execution path governance.

Tokyo, Japan

Jun. 2026 - Aug. 2026

Waseda University

VISITING STUDENT

- Conducted research under the supervision of Prof. Shigeru Fujimura, applying machine learning techniques to automatically detect printing defects and illegible handwriting on student ID cards.

Tokyo, Japan

Jul. 2018 - Aug. 2018

Publications

CONFERENCES

[1] Membership inference attacks by exploiting loss trajectory

Yiyong Liu, Zhengyu Zhao, Michael Backes, Yang Zhang

In ACM SIGSAC Conference on Computer and Communications Security (CCS 2022).

[2] Auditing Membership Leakages of Multi-Exit Networks

Zheng Li, Yiyong Liu, Xinlei He, Ning Yu, Michael Backes, Yang Zhang

In ACM SIGSAC Conference on Computer and Communications Security (CCS 2022).

[3] SoK: Data Reconstruction Attacks Against Machine Learning Models: Definition, Metrics, and Benchmark

Rui Wen*, Yiyong Liu*, Michael Backes, Yang Zhang (*Equal Contribution)

PREPRINTS

[1] No Free Efficiency: Revisiting the Trade-off Between Training Efficiency and Model Vulnerability

Yiyong Liu, Jun Sakuma, Michael Backes, Rui Wen

arXiv, 2026.

[2] Near-Duplicate Families Break Exact-Record Membership Inference

Yiyong Liu, Jiayang Liu, Yixin Tan, Lu Sun, Rui Wen

arXiv, 2026.

[3] Capability-Routed Guard: Defending Large Reasoning Models Against Reasoning-Centric Jailbreaks

Yiyong Liu, Yixin Wu, Jun Sakuma

arXiv, 2026.

[4] Trust Me, Import This: Dependency Steering Attacks via Malicious Agent Skills

Yiyong Liu*, Chia-Yi Hsu*, Chun-Ying Huang, Michael Backes, Rui Wen, Chia-Mu Yu (*Equal Contribution)

arXiv, 2026.

[5] Efficient data-free model stealing with label diversity

Yiyong Liu, Rui Wen, Michael Backes, Yang Zhang

arXiv, 2024.

[6] Transferable availability poisoning attacks

Yiyong Liu, Michael Backes, Xiao Zhang

arXiv, 2023.

Skills

Programming	Python, C/C++
Machine Learning	PyTorch and the Hugging Face ecosystem; training, fine-tuning, and evaluation of large language models
Infrastructure	Linux, Docker, and GPU cluster environments (Slurm, HPC)

Professional Service

Reviewer / External Reviewer: PETS 2027, CVPR 2026, NeurIPS 2025, KDD 2025, ACML 2024, NDSS 2023, USENIX Security 2022, PoPETs 2022, AsiaCCS 2022, AAAI 2022, ACM CCS 2021