Changsheng Wang

Email: wangc168@msu.edu | Phone: 517 204 5931 | Google Schorlar

Research Focus

Trustworthy ML: Machine unlearning, adversarial attack and defense

Generative AI: Diffusion models, large language models (LLMs), multi-modal models, jailbreaking attacks, watermarking

Education

Michigan State University (MSU) Sept 2024 – present

Ph.D. in Computer Science

Advisor: Prof. Sijia Liu

University of Science and Technology of China (USTC) Sept 2020 – Apr 2024

B.S. in Data Science and Big Data Technology (Rank 5/35)

Working Experience

Research Assistant, The OPTML Lab @ Michigan State University Sept 2024 – present

- **PhD Advisor:** Prof. Sijia Liu
- Advancing LLM unlearning across the full optimization-model-data stack

Research Intern, University of Wisconsin–Madison Mar 2023 – Mar 2024

- **Mentor:** Prof. Chaowei Xiao
- Focused on addressing adversarial vulnerabilities in diffusion models and vision-language models

Remote Research Intern, National University of Singapore Sept 2022 - Mar 2023

- **Mentor:** Prof. Wenjie Wang
- Worked to ensure the security and robustness of recommender systems

Research Projects

Machine Unlearning for LLMs @ MSU Sept 2024 – present

- Enhancing the robustness of LLM unlearning against relearning and fine-tuning attacks through unlearning invariance with invariant risk minimization (IRM) (manuscript in preparation for ICML 2025)
- Exploring the influence of LLM watermarking on LLM unlearning from a data-centric perspective (manuscript in preparation for ICML 2025)

Adversarial Robustness of Generative Models @ UW-Madison Mar 2023 – Mar 2024

- Safeguarding vision-language models (VLMs) against patched visual prompt injectors
- Reduced the attack success rate to 0%–5% on two leading VLMs while achieving 67.3%–95.0% context recovery for benign images (paper under review; arXiv version [P1])

Recommender System Security @ NUS Sept 2022 - Mar 2023

- Proposed a novel attack paradigm, Target User Attack, for recommender systems (Publication @ WWW'24, [P2])
- Created a security toolbox for recommender systems, combining datasets, standardized code, hyperparameter settings, logs, attack strategies, budgets, and evaluation metrics. (Publication @ RecSys'23, [P3])

Publications

[P1] Safeguarding Vision-Language Models Against Patched Visual Prompt Injectors (arXiv)
Jiachen Sun, ***Changsheng Wang***, Jiongxiao Wang, Yiwei Zhang, Chaowei Xiao

[P2] Uplift Modeling for Target User Attacks on Recommender Systems. *WWW' 2024*, *Oral*

Wenjie Wang*, ***Changsheng Wang***, Fuli Feng, Daizong Ding, Tat-Seng Chua (*Equal contribution)

[P3] RecAD: Towards A Unified Library for Recommender Attack and Defense. RecSys' 2023

Changsheng Wang, Jianbai Ye, Wenjie Wang, Chongming Gao, Fuli Feng, Xiangnan He

Services

Student Activity Chair: The 3rd New Frontiers in Adversarial Machine Learning (AdvML Frontiers @ NeurIPS2024)

Conference Reviewer: NeurIPS'2024, ICLR'2024

Honors and Awards

USTC Outstanding Graduation Thesis (Top 4%) 2024

USTC Outstanding Students Award (Top 2%) 2021 – 2023

USTC Outstanding Freshman Scholarship (Top 4%) 2020

USTC Robot Game Second Prize (2/16 teams) 2023

Skills

Languages: Python, C, C++, SQL, Matlab, JavaScript

Frameworks: PyTorch, TensorFlow, Keras, Django, Flask

LLM/VLM: LLaMA, Mixtral, LLaVA, Flamingo, miniGPT4

Tools: LaTeX, MySQL, Docker, Git