Yang Zhang | CV

zhang@cispa.de • yangzhangalmo.github.io

last update: November 5, 2026

Employment

CISPA Helmholtz Center for Information Security — **Saarbrücken, Germany**
Tenured Faculty, equivalent to full professor — *5/2023 -*

CISPA Helmholtz Center for Information Security — **Saarbrücken, Germany**
Tenure-Track Faculty — *2/2020 - 4/2023*

CISPA Helmholtz Center for Information Security — **Saarbrücken, Germany**
Research Group Leader — *1/2019 - 1/2020*

CISPA, Saarland University — **Saarbrücken, Germany**
Postdoctoral Researcher — *1/2017 - 12/2018*

Education

University of Luxembourg — **Luxembourg, Luxembourg**
Ph.D. in Computer Science, highest honor — *12/2012 - 11/2016*

Shandong University — **Jinan, China**
Master in Computer Science — *9/2009 - 6/2012*

University of Luxembourg — **Luxembourg, Luxembourg**
Master in Informatics, exchange student — *9/2010 - 10/2011*

Shandong University — **Jinan, China**
Bachelor in Software Engineering — *9/2005 - 6/2009*

Research Interests

- Trustworthy AI (Safety, Privacy, and Security)
- Misinformation, Hate Speech, and Memes
- Social Network Analysis

Awards

- "Busy Beaver" Teaching Award for Seminar "AI Safety" at Saarland University (2025 Winter)
- EMNLP 2026 Award Nomination
- AI 2000 Most Influential Scholar Award Honorable Mention (Security and Privacy) 2026
- CVPR 2026 Highlight
- AAAI 2026 Oral Presentation
- AI 2000 Most Influential Scholar Award Honorable Mention (Security and Privacy) 2025
- Best Machine Learning and Security Paper in Cybersecurity Award 2025
- ACL 2025 Oral Presentation
- Best Paper Finalist at CSAW Europe 2024

- Best Paper Finalist at CSAW Europe 2023
- ICLR 2023 Spotlight Paper
- AAAI 2023 Oral Presentation
- Best Paper Award Honorable Mention at CCS 2022
- "Busy Beaver" Teaching Award Nomination for Seminar "Privacy of Machine Learning" at Saarland University (2022 Winter)
- "Busy Beaver" Teaching Award Nomination for Advanced Lecture "Machine Learning Privacy" at Saarland University (2022 Summer)
- "Busy Beaver" Teaching Award for Seminar "Privacy of Machine Learning" at Saarland University (2021 Winter)
- Distinguished Paper Award at NDSS 2019

Publication

My publication list can also be found at DBLP and Google Scholar; however, they may not be up to date.

Conference

[1] Zeyuan Chen and Ye Leng and Chi Cui and Mengfei Liang and Bo Shao and Xinyu Zhang and Yun Shen and Michael Backes and Xinyue Shen and **Yang Zhang**. Can We Trust LLM Security Research? A Reproducibility Study of Large Language Model Security Papers in Tier-1 Security Conference. In *IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2027.

[2] Xinyu Zhang and Ziqing Yang and Michael Backes and **Yang Zhang**. PersonaHate: A Scalable Persona-Based Data Synthesis Pipeline for Hate Speech Analysis. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2026.

[3] Yage Zhang and Yukun Jiang and Zeyuan Chen and Michael Backes and Xinyue Shen and **Yang Zhang**. Real Money, Fake Models: Deceptive Model Claims in Shadow APIs. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2026.

[4] Bo Shao and Xinyue Shen and **Yang Zhang**. Does AI Amplify Online Hate? A Multi-Modal Measurement Study on 4chan's /pol/. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2026.

[5] Yage Zhang and Xinyue Shen and Yukun Jiang and Michael Backes and **Yang Zhang**. "Shut Up and Let Me Enjoy My Otome": Understanding and Measuring the Toxicity in Otome Game Communities. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2026.

[6] Chia-Yi Hsu and Yu-Lin Tsai and Yu Zhe and Yan-Lun Chen and Chih-Hsun Lin and Chia-Mu Yu and **Yang Zhang** and Chun-Ying Huang and Jun Sakuma. BadTV: Unveiling Backdoor Threats in Third-Party Task Vectors. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2026.

[7] Junjie Chu and Mingjie Li and Ziqing Yang and Ye Leng and Chenhao Lin and Chao Shen and Michael Backes and Yun Shen and **Yang Zhang**. JADES: A Universal Framework for Jailbreak Assessment via Decompositional Scoring. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 2026.

[8] Yicong Tan and Yun Shen and **Yang Zhang**. He Said, She Said, They Simulated: Simulating Situated Gossip with LLM Agents. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 2026.

[9] Yichuan Mo and Yukun Jiang and Yanbo Shi and Mingjie Li and Michael Backes and **Yang Zhang** and Yisen Wang. TrustDLM: Benchmarking Trustworthiness in Diffusion Language Models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 2026.

[10] Junjie Chu and Mingjie Li and Zeyuan Chen and Ye Leng and Yisen Wang and Michael Backes and **Yang Zhang**. Behavior Carries Over: An AI Sandbagging Detection Method Inspired by Honest Inertia. In *Findings of the Association for Computational Linguistics: EMNLP (EMNLP Findings)*. ACL, 2026.

[11] Ye Leng and Junjie Chu and Mingjie Li and Michael Backes and Yun Shen and **Yang Zhang**. RW2Unsafe: Towards More Accurate Real-World Unsafe Generation Evaluation for T2I Models. In *Findings of the Association for Computational Linguistics: EMNLP (EMNLP Findings)*. ACL, 2026.

[12] Yicong Tan and **Yang Zhang**. Query-Efficient Memory Injection Attacks on LLM Agents. In *Findings of the Association for Computational Linguistics: EMNLP (EMNLP Findings)*. ACL, 2026.

[13] Ziqing Yang and Yixin Wu and Yun Shen and Wei Dai and Michael Backes and **Yang Zhang**. The Challenge of Identifying the Origin of Black-Box Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP (EMNLP Findings)*. ACL, 2026.

[14] Tianze Chang and Junjie Chu and Yixin Wu and Michael Backes and **Yang Zhang**. GuardrailAgent: An Agentic Framework for Jailbreak Defense. In *Findings of the Association for Computational Linguistics: EMNLP (EMNLP Findings)*. ACL, 2026.

[15] Mingjie Li and Michael Backes and **Yang Zhang** and Yisen Wang. Long-Short Memory Steering to Mitigate Unfaithful Self-Brainwashing. In *Findings of the Association for Computational Linguistics: EMNLP (EMNLP Findings)*. ACL, 2026.

[16] Chi Cui and Yixin Wu and Michael Backes and **Yang Zhang**. Beyond Text-Intrinsic Features: An Agentic Framework for Evidence-Based AI-Generated Text Detection. In *Findings of the Association for Computational Linguistics: EMNLP (EMNLP Findings)*. ACL, 2026.

[17] Bo Shao and Yihan Ma and Yiting Qu and Michael Backes and **Yang Zhang**. MemeTrace: A Multimodal LLM-Based Agent for Safety-Aware Meme Provenance Analysis. In *Findings of the Association for Computational Linguistics: EMNLP (EMNLP Findings)*. ACL, 2026.

[18] Xinyu Zhang and Yixin Wu and Boyang Zhang and Chenhao Lin and Chao Shen and Michael Backes and **Yang Zhang**. GEO-Detective: Unveiling Location Privacy Risks in Images with LLM Agents. In *European Conference on Computer Vision (ECCV)*, pages 547–565. Springer, 2026.

[19] Quan Zhao and Linkang Du and Yuntao Wang and Zhou Su and Zheng Li and Xiangshan Gao and **Yang Zhang**. Unveiling the Pitfalls of Data-Free Backdoor Detection Against Pre-Trained Models. In *USENIX Security Symposium (USENIX Security)*, pages 261–280. USENIX, 2025.

[20] Yukun Jiang and Hai Huang and Mingjie Li and Yage Zhang and Michael Backes and **Yang Zhang**. Sparse Models, Sparse Safety: Unsafe Routes in Mixture-of-Experts LLMs. In *International Conference on Machine Learning (ICML)*. PMLR, 2026.

[21] Isabella Duan and Xudong Pan and Yawen Duan and Adam Gleave and Ranjie Duan and **Yang Zhang** and Xiaojian Li and Chaochao Lu and Naying Hu and Sören Mindermann and Dongrui Liu and Jie Fu and Peng Xu and Tianxing He and Xudong Guo and Chen Zheng and Wenqi Chen and Jianfeng Cao and Geng Hong and Jiarun Dai and Yinpeng Dong and Brian Tse and Xia Hu and Min Yang. Position: Preparing for AI Systems That Deceive Developers. In *International Conference on Machine Learning (ICML Position Paper Track)*. PMLR, 2026.

[22] Wai Man Si and Mingjie Li and Michael Backes and **Yang Zhang**. Pruning Unsafe Tickets: A Resource-Efficient Framework for Safer and More Robust LLMs. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 26285–26302. ACL, 2026.

[23] Yukun Jiang and Xinyue Shen and Michael Backes and Zheng Li and **Yang Zhang**. Open Schrödinger's Closed Box: Identifying Retrieval Augmented Generation in API-Accessible Large Language Model Services. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4558–4580. ACL, 2026.

[24] Yage Zhang and Yukun Jiang and Michael Backes and **Yang Zhang**. DE-CLIP: Few-Shot Anomaly Detection via Difference-Guided Embedding Editing. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2396–2407. ACL, 2026.

[25] Rui Zhang and Hongwei Li and Yun Shen and Xinyue Shen and Wenbo Jiang and Guowen Xu and Yang Liu and Michael Backes and **Yang Zhang**. The Art of (Mis)alignment: How Fine-Tuning Methods Effectively Misalign and Realign LLMs in Post-Training. In *Findings of the Association for Computational Linguistics: ACL (ACL Findings)*, pages 3329–3354. ACL, 2026.

[26] Ziqing Yang and Yixin Wu and Rui Wen and Michael Backes and **Yang Zhang**. Peering Behind the Shield: Guardrail Identification in Large Language Models. In *Findings of the Association for Computational Linguistics: ACL (ACL Findings)*, pages 11663–11676. ACL, 2026.

[27] Yixin Wu and Rui Wen and Chi Cui and Michael Backes and **Yang Zhang**. InferPilot: Autonomous Inference Attacks Against ML Services With LLM-Based Agents. In *Findings of the Association for Computational Linguistics: ACL (ACL Findings)*, pages 11781–11801. ACL, 2026.

[28] Zeyuan Chen and Ziqing Yang and Yihan Ma and Michael Backes and **Yang Zhang**. PeerCheck: Enhancing LLM-Generated Academic Reviews Towards Human-Level Quality. In *Findings of the Association for Computational Linguistics: ACL (ACL Findings)*, pages 23362–23386. ACL, 2026.

[29] Chi Cui and Yixin Wu and Michael Backes and **Yang Zhang**. Rethinking Assessments of Prompt Injection Attacks. In *Findings of the Association for Computational Linguistics: ACL (ACL Findings)*, pages 23773–23799. ACL, 2026.

[30] Mingjie Li and Wai Man Si and Michael Backes and **Yang Zhang**. Reward Yourself: Efficient Self Rewards for Trustworthy Sampling. In *Findings of the Association for Computational Linguistics: ACL (ACL Findings)*, pages 31070–31086. ACL, 2026.

[31] Ye Leng and Junjie Chu and Mingjie Li and Chenhao Lin and Chao Shen and Michael Backes and Yun Shen and **Yang Zhang**. When Understanding Becomes a Risk: Authenticity and Safety Risks in the Emerging Image Generation Paradigm. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 39372–39382. IEEE, 2026.

[32] Boyang Zhang and Istemi Ekin Akkus and Ruichuan Chen and Alice Dethise and Klaus Satzke and Ivica Rimac and **Yang Zhang**. Defeating Cerberus: Privacy-Leakage Mitigation in Vision Language Models. In *Findings of the Association for Computational Linguistics: EACL (EACL Findings)*, pages 2952–2965. ACL, 2026.

[33] Hanwei Zhang and Luo Cheng and Rui Wen and **Yang Zhang** and Lijun Zhang and Holger Hermanns. SL-CBM: Enhancing Concept Bottleneck Models with Semantic Locality for Better Interpretability. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 38093–38101. AAAI, 2026.

[34] Yukun Jiang and Mingjie Li and Michael Backes and **Yang Zhang**. Adjacent Words, Divergent Intents: Jailbreaking Large Language Models via Task Concurrency. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2025.

[35] Mingjie Li and Wai Man Si and Michael Backes and **Yang Zhang** and Yisen Wang. Finding and Reactivating Post-Trained LLMs' Hidden Safety Mechanisms. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2025.

[36] Boyang Zhang and Yicong Tan and Yun Shen and Ahmed Salem and Michael Backes and Savvas Zannettou and **Yang Zhang**. Breaking Agents: Compromising Autonomous LLM Agents Through Malfunction Amplification. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 34964–34976. ACL, 2025.

[37] Yiting Qu and Ziqing Yang and Yihan Ma and Michael Backes and Savvas Zannettou and **Yang Zhang**. Hate in Plain Sight: On the Risks of Moderating AI-Generated Hateful Illusions. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2025.

[38] Yiting Qu and Xinyue Shen and Yixin Wu and Michael Backes and 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)*, pages 3221–3235. ACM, 2025.

[39] Yiting Qu and Michael Backes and **Yang Zhang**. Bridging the Gap in Vision Language Models in Identifying Unsafe Concepts Across Modalities. In *USENIX Security Symposium (USENIX Security)*, pages 957–976. USENIX, 2025.

[40] Rui Wen and Yiyong Liu and Michael Backes and **Yang Zhang**. SoK: Data Reconstruction Attacks Against Machine Learning Models: Definition, Metrics, and Benchmark. In *USENIX Security Symposium (USENIX Security)*, pages 5601–5620. USENIX, 2025.

[41] Xinyue Shen and Yixin Wu and Yiting Qu and Michael Backes and Savvas Zannettou and **Yang Zhang**. HateBench: Benchmarking Hate Speech Detectors on LLM-Generated Content and Hate Campaigns. In *USENIX Security Symposium (USENIX Security)*, pages 221–240. USENIX, 2025.

[42] Yihan Ma and Xinyue Shen and Yiting Qu and Ning Yu and Michael Backes and 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)*, pages 4703–4722. USENIX, 2025.

[43] Yixin Wu and Ning Yu and Michael Backes and Yun Shen and **Yang Zhang**. On the Proactive Generation of Unsafe Images From Text-To-Image Models Using Benign Prompts. In *USENIX Security Symposium (USENIX Security)*, pages 917–936. USENIX, 2025.

[44] Yixin Wu and Ziqing Yang and Yun Shen and Michael Backes and **Yang Zhang**. Synthetic Artifact Auditing: Tracing LLM-Generated Synthetic Data Usage in Downstream Applications. In *USENIX Security Symposium (USENIX Security)*, pages 1689–1708. USENIX, 2025.

[45] Dayong Ye and Tianqing Zhu and Shang Wang and Bo Liu and Leo Yu Zhang and Wanlei Zhou and **Yang Zhang**. Data-Free Model-Related Attacks: Unleashing the Potential of Generative AI. In *USENIX Security Symposium (USENIX Security)*, pages 1709–1727. USENIX, 2025.

[46] Dayong Ye and Tianqing Zhu and Jiayang Li and Kun Gao and Bo Liu and Leo Yu Zhang and Wanlei Zhou and **Yang Zhang**. Data Duplication: A Novel Multi-Purpose Attack Paradigm in Machine Unlearning. In *USENIX Security Symposium (USENIX Security)*, pages 6399–6418. USENIX, 2025.

[47] Hao Li and Zheng Li and Siyuan Wu and Yutong Ye and Min Zhang and Dengguo Feng and **Yang Zhang**. Enhanced Label-Only Membership Inference Attacks with Fewer Queries. In *USENIX Security Symposium (USENIX Security)*, pages 5465–5483. USENIX, 2025.

[48] Yuke Hu and Zheng Li and Zhihao Liu and **Yang Zhang** and Zhan Qin and Kui Ren and Chun Chen. Membership Inference Attacks Against Vision-Language Models. In *USENIX Security Symposium (USENIX Security)*, pages 1589–1608. USENIX, 2025.

[49] Atilla Akkus and Masoud Poorghaffar Aghdam and Mingjie Li and Junjie Chu and Michael Backes and **Yang Zhang** and Sinem Sav. Generated Data with Fake Privacy: Hidden Dangers of Fine-tuning Large Language Models on Generated Data. In *USENIX Security Symposium (USENIX Security)*, pages 8075–8093. USENIX, 2025.

[50] Yule Liu and Zhiyuan Zhong and Yifan Liao and Zhen Sun and Jingyi Zheng and Jiaheng Wei and Qingyuan Gong and Fenghua Tong and Yang Chen and **Yang Zhang** and Xinlei He. On the Generalization Ability of Machine-Generated Text Detectors. In *ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 5674–5685. ACM, 2025.

[51] Junjie Chu and Yugeng Liu and Ziqing Yang and Xinyue Shen and Michael Backes and **Yang Zhang**. JailbreakRadar: Comprehensive Assessment of Jailbreak Attacks Against LLMs. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 21538–21566. ACL, 2025.

[52] Xinyue Shen and Yun Shen and Michael Backes and **Yang Zhang**. When GPT Spills the Tea: Comprehensive Assessment of Knowledge File Leakage in GPTs. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 19096–19111. ACL, 2025.

[53] Zhen Sun and Zongmin Zhang and Xinyue Shen and Ziyi Zhang and Yule Liu and Michael Backes and **Yang Zhang** and Xinlei He. Are We in the AI-Generated Text World Already? Quantifying and Monitoring AIGT on Social Media. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 22975–23005. ACL, 2025.

[54] Rui Zhang and Yun Shen and Hongwei Li and Wenbo Jiang and Hanxiao Chen and Yuan Zhang and Guowen Xu and **Yang Zhang**. The Ripple Effect: On Unforeseen Complications of Backdoor Attacks. In *International Conference on Machine Learning (ICML)*. PMLR, 2025.

[55] Junjie Chu and Yugeng Liu and Xinlei He and Michael Backes and **Yang Zhang** and Ahmed Salem. Neeko: Model Hijacking Attacks Against Generative Adversarial Networks. In *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2025.

[56] Xinyue Shen and Yun Shen and Michael Backes and **Yang Zhang**. GPTracker: A Large-Scale Measurement of Misused GPTs. In *IEEE Symposium on Security and Privacy (S&P)*, pages 317–335. IEEE, 2025.

[57] Yicong Tan and Xinyue Shen and Yun Shen and Michael Backes and **Yang Zhang**. On the Effectiveness of Prompt Stealing Attacks on In-The-Wild Prompts. In *IEEE Symposium on Security and Privacy (S&P)*, pages 355–373. IEEE, 2025.

[58] Mingjie Li and Wai Man Si and Michael Backes and **Yang Zhang** and Yisen Wang. SaLoRA: Safety-Alignment Preserved Low-Rank Adaptation. In *International Conference on Learning Representations (ICLR)*, 2025.

[59] Yan Pang and Aiping Xiong and **Yang Zhang** and Tianhao Wang. Towards Understanding Unsafe Video Generation. In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2025.

[60] Rui Wen and Michael Backes and **Yang Zhang**. Understanding Data Importance in Machine Learning Attacks: Does Valuable Data Pose Greater Harm? In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2025.

[61] Yihan Ma and Xinyue Shen and Yixin Wu and Boyang Zhang and Michael Backes and **Yang Zhang**. The Death and Life of Great Prompts: Analyzing the Evolution of LLM Prompts from the Structural Perspective. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 21990–22001. ACL, 2024.

[62] Yukun Jiang and Zheng Li and Xinyue Shen and Yugeng Liu and Michael Backes and **Yang Zhang**. ModScan: Measuring Stereotypical Bias in Large Vision-Language Models from Vision and Language Modalities. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12814–12845. ACL, 2024.

[63] Junjie Chu and Zeyang Sha and Michael Backes and **Yang Zhang**. Reconstruct Your Previous Conversations! Comprehensively Investigating Privacy Leakage Risks in Conversations with GPT Models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6584–6600. ACL, 2024.

[64] Rui Wen and Zheng Li and Michael Backes and **Yang Zhang**. Membership Inference Attacks Against In-Context Learning. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 3481–3495. ACM, 2024.

[65] Yixin Wu and Yun Shen and Michael Backes and **Yang Zhang**. Image-Perfect Imperfections: Safety, Bias, and Authenticity in the Shadow of Text-To-Image Model Evolution. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 4837–4851. ACM, 2024.

[66] Jinghuai Zhang and Jianfeng Chi and Zheng Li and Kunlin Cai and **Yang Zhang** and Yuan Tian. BadMerging: Backdoor Attacks Against Model Merging. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 4450–4464. ACM, 2024.

[67] Zeyang Sha and Yicong Tan and Mingle Li and Michael Backes and **Yang Zhang**. ZeroFake: Zero-Shot Detection of Fake Images Generated and Edited by Text-to-Image Generation Models. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 4852–4866. ACM, 2024.

[68] Hao Li and Zheng Li and Siyuan Wu and Chengrui Hu and Yutong Ye and Min Zhang and Dengguo Feng and **Yang Zhang**. SeqMIA: Sequential-Metric Based Membership Inference Attack. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 3496–3510. ACM, 2024.

[69] Xinlei He and Xinyue Shen and Zeyuan Chen and Michael Backes and **Yang Zhang**. MGTBench: Benchmarking Machine-Generated Text Detection. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 2251–2265. ACM, 2024.

[70] Xinyue Shen and Zeyuan Chen and Michael Backes and Yun Shen and **Yang Zhang**. "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1671–1685. ACM, 2024.

[71] Rui Zhang and Hongwei Li and Rui Wen and Wenbo Jiang and Yuan Zhang and Michael Backes and Yun Shen and **Yang Zhang**. Instruction Backdoor Attacks Against Cutomized LLMs. In *USENIX Security Symposium (USENIX Security)*. USENIX, 2024.

[72] Xinyue Shen and Yiting Qu and Michael Backes and **Yang Zhang**. Prompt Stealing Attacks Against Text-to-Image Generation Models. In *USENIX Security Symposium (USENIX Security)*. USENIX, 2024.

[73] Boyang Zhang and Zheng Li and Ziqing Yang and Xinlei He and Michael Backes and Mario Fritz and **Yang Zhang**. SecurityNet: Assessing Machine Learning Vulnerabilities on Public Models. In *USENIX Security Symposium (USENIX Security)*. USENIX, 2024.

[74] Yixin Wu and Rui Wen and Michael Backes and Pascal Berrang and Mathias Humbert and Yun Shen and **Yang Zhang**. Quantifying Privacy Risks of Prompts in Visual Prompt Learning. In *USENIX Security Symposium (USENIX Security)*. USENIX, 2024.

[75] Hai Huang and Zhengyu Zhao and Michael Backes and Yun Shen and **Yang Zhang**. Composite Backdoor Attacks Against Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL (NAACL Findings)*. ACL, 2024.

[76] Yukun Jiang and Xinyue Shen and Rui Wen and Zeyang Sha and Junjie Chu and Yugeng Liu and Michael Backes and **Yang Zhang**. Games and Beyond: Analyzing the Bullet Chats of Esports Livestreaming. In *International Conference on Weblogs and Social Media (ICWSM)*, pages 761–773. AAAI, 2024.

[77] Yiting Qu and Zhikun Zhang and Yun Shen and Michael Backes and **Yang Zhang**. FAKEPCD: Fake Point Cloud Detection via Source Attribution. In *ACM Asia Conference on Computer and Communications Security (ASIACCS)*, pages 930–946. ACM, 2024.

[78] Xinlei He and Savvas Zannettou and Yun Shen and **Yang Zhang**. You Only Prompt Once: On the Capabilities of Prompt Learning on Large Language Models to Tackle Toxic Content. In *IEEE Symposium on Security and Privacy (S&P)*, pages 770–787. IEEE, 2024.

[79] Tianshuo Cong and Xinlei He and Yun Shen and **Yang Zhang**. Test-Time Poisoning Attacks Against Test-Time Adaptation Models. In *IEEE Symposium on Security and Privacy (S&P)*, pages 1306–1324. IEEE, 2024.

[80] Minxing Zhang and Ning Yu and Rui Wen and Michael Backes and **Yang Zhang**. Generated Distributions Are All You Need for Membership Inference Attacks Against Generative Models. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2024.

[81] Zeyang Sha and Zheng Li and Ning Yu and **Yang Zhang**. DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Generation Models. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 3418–3432. ACM, 2023.

[82] Yiting Qu and Xinyue Shen and Xinlei He and Michael Backes and 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)*, pages 3403–3417. ACM, 2023.

[83] Joann Qiongna Chen and Tianhao Wang and Zhikun Zhang and **Yang Zhang** and Somesh Jha and Zhou Li. Differentially Private Resource Allocation. In *Annual Computer Security Applications Conference (ACSAC)*, pages 772–786. ACM, 2023.

[84] Boyang Zhang and Xinlei He and Yun Shen and Tianhao Wang and **Yang Zhang**. A Plot is Worth a Thousand Words: Model Information Stealing Attacks via Scientific Plots. In *USENIX Security Symposium (USENIX Security)*, pages 5289–5306. USENIX, 2023.

[85] Wai Man Si and Michael Backes and **Yang Zhang** and Ahmed Salem. Two-in-One: A Model Hijacking Attack Against Text Generation Models. In *USENIX Security Symposium (USENIX Security)*, pages 2223–2240. USENIX, 2023.

[86] Zheng Li and Ning Yu and Ahmed Salem and Michael Backes and Mario Fritz and **Yang Zhang**. UnGANable: Defending Against GAN-based Face Manipulation. In *USENIX Security Symposium (USENIX Security)*, pages 7213–7230. USENIX, 2023.

[87] Min Chen and Zhikun Zhang and Michael Backes and Tianhao Wang and **Yang Zhang**. FACE-AUDITOR: Data Auditing in Facial Recognition Systems. In *USENIX Security Symposium (USENIX Security)*, pages 7195–7212. USENIX, 2023.

[88] Haiming Wang and Zhikun Zhang and Tianhao Wang and Shibo He and Michael Backes and Jiming Chen and **Yang Zhang**. PrivTrace: Differentially Private Trajectory Synthesis by Adaptive Markov Model. In *USENIX Security Symposium (USENIX Security)*, pages 1649–1666. USENIX, 2023.

[89] Yihan Ma and Zhikun Zhang and Ning Yu and Xinlei He and Michael Backes and Yun Shen and **Yang Zhang**. Generated Graph Detection. In *International Conference on Machine Learning (ICML)*, pages 23412–23428. PMLR, 2023.

[90] Ziqing Yang and Xinlei He and Zheng Li and Michael Backes and Mathias Humbert and Pascal Berrang and **Yang Zhang**. Data Poisoning Attacks Against Multimodal Encoders. In *International Conference on Machine Learning (ICML)*, pages 39299–39313. PMLR, 2023.

[91] Kai Mei and Zheng Li and Zhenting Wang and **Yang Zhang** and Shiqing Ma. NOTABLE: Transferable Backdoor Attacks Against Prompt-based NLP Models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 15551–15565. ACL, 2023.

[92] Zeyang Sha and Xinlei He and Ning Yu and Michael Backes and **Yang Zhang**. Can't Steal? Cont-Steal! Contrastive Stealing Attacks Against Image Encoders. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16373–16383. IEEE, 2023.

[93] 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)*, pages 293–310. IEEE, 2023.

[94] Rui Wen and Zhengyu Zhao and Zhuoran Liu and Michael Backes and Tianhao Wang and **Yang Zhang**. Is Adversarial Training Really a Silver Bullet for Mitigating Data Poisoning? In *International Conference on Learning Representations (ICLR)*, 2023.

[95] Yugeng Liu and Zheng Li and Michael Backes and Yun Shen and **Yang Zhang**. Backdoor Attacks Against Dataset Distillation. In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2023.

[96] Xiaojian Yuan and Kejiang Chen and Jie Zhang and Weiming Zhang and Nenghai Yu and **Yang Zhang**. Pseudo Label-Guided Model Inversion Attack via Conditional Generative Adversarial Network. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 3349–3357. AAAI, 2023.

[97] Yufei Chen and Chao Shen and Yun Shen and Cong Wang and **Yang Zhang**. Amplifying Membership Exposure via Data Poisoning. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2022.

[98] Wai Man Si and Michael Backes and Jeremy Blackburn and Emiliano De Cristofaro and Gianluca Stringhini and Savvas Zannettou and **Yang Zhang**. Why So Toxic?: Measuring and Triggering Toxic Behavior in Open-Domain Chatbots. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 2659–2673. ACM, 2022.

[99] Hai Huang and Zhikun Zhang and Yun Shen and Michael Backes and Qi Li and **Yang Zhang**. On the Privacy Risks of Cell-Based NAS Architectures. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1427–1441. ACM, 2022.

[100] Yiyong Liu and Zhengyu Zhao and Michael Backes and **Yang Zhang**. Membership Inference Attacks by Exploiting Loss Trajectory. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 2085–2098. ACM, 2022.

[101] Zheng Li and Yiyong Liu and Xinlei He and Ning Yu and Michael Backes and **Yang Zhang**. Auditing Membership Leakages of Multi-Exit Networks. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1917–1931. ACM, 2022.

[102] Min Chen and Zhikun Zhang and Tianhao Wang and Michael Backes and Mathias Humbert and **Yang Zhang**. Graph Unlearning. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 499–513. ACM, 2022.

[103] Tianshuo Cong and Xinlei He and **Yang Zhang**. SSLGuard: A Watermarking Scheme for Self-supervised Learning Pre-trained Encoders. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 579–593. ACM, 2022.

[104] Yun Shen and Yufei Han and Zhikun Zhang and Min Chen and Ting Yu and Michael Backes and **Yang Zhang** and Gianluca Stringhini. Finding MNEMON: Reviving Memories of Node Embeddings. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 2643–2657. ACM, 2022.

[105] Xinlei He and Hongbin Liu and Neil Zhenqiang Gong and **Yang Zhang**. Semi-Leak: Membership Inference Attacks Against Semi-supervised Learning. In *European Conference on Computer Vision (ECCV)*, pages 365–381. Springer, 2022.

[106] Yugeng Liu and Rui Wen and Xinlei He and Ahmed Salem and Zhikun Zhang and Michael Backes and Emiliano De Cristofaro and Mario Fritz and **Yang Zhang**. ML-Doctor: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models. In *USENIX Security Symposium (USENIX Security)*, pages 4525–4542. USENIX, 2022.

[107] Yufei Chen and Chao Shen and Cong Wang and **Yang Zhang**. Teacher Model Fingerprinting Attacks Against Transfer Learning. In *USENIX Security Symposium (USENIX Security)*, pages 3593–3610. USENIX, 2022.

[108] Zhikun Zhang and Min Chen and Michael Backes and Yun Shen and **Yang Zhang**. Inference Attacks Against Graph Neural Networks. In *USENIX Security Symposium (USENIX Security)*, pages 4543–4560. USENIX, 2022.

[109] Xinyue Shen and Xinlei He and Michael Backes and Jeremy Blackburn and Savvas Zannettou and **Yang Zhang**. On Xing Tian and the Perseverance of Anti-China Sentiment Online. In *International Conference on Weblogs and Social Media (ICWSM)*, pages 944–955. AAAI, 2022.

[110] Yun Shen and Xinlei He and Yufei Han and **Yang Zhang**. Model Stealing Attacks Against Inductive Graph Neural Networks. In *IEEE Symposium on Security and Privacy (S&P)*, pages 1175–1192. IEEE, 2022.

[111] Ahmed Salem and Rui Wen and Michael Backes and Shiqing Ma and **Yang Zhang**. Dynamic Backdoor Attacks Against Machine Learning Models. In *IEEE European Symposium on Security and Privacy (Euro S&P)*, pages 703–718. IEEE, 2022.

[112] Ahmed Salem and Michael Backes and **Yang Zhang**. Get a Model! Model Hijacking Attack Against Machine Learning Models. In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2022.

[113] Junhao Zhou and Yufei Chen and Chao Shen and **Yang Zhang**. Property Inference Attacks Against GANs. In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2022.

[114] Xinlei He and **Yang Zhang**. Quantifying and Mitigating Privacy Risks of Contrastive Learning. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 845–863. ACM, 2021.

[115] Min Chen and Zhikun Zhang and Tianhao Wang and Michael Backes and Mathias Humbert and **Yang Zhang**. When Machine Unlearning Jeopardizes Privacy. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 896–911. ACM, 2021.

[116] Minxing Zhang and Zhaochun Ren and Zihan Wang and Pengjie Ren and Zhumin Chen and Pengfei Hu and **Yang Zhang**. Membership Inference Attacks Against Recommender Systems. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 864–879. ACM, 2021.

[117] Zheng Li and **Yang Zhang**. Membership Leakage in Label-Only Exposures. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 880–895. ACM, 2021.

[118] Xiaoyi Chen and Ahmed Salem and Michael Backes and Shiqing Ma and Qingni Shen and Zhonghai Wu and **Yang Zhang**. BadNL: Backdoor Attacks Against NLP Models with Semantic-preserving Improvements. In *Annual Computer Security Applications Conference (ACSAC)*, pages 554–569. ACSAC, 2021.

[119] Xinlei He and Jinyuan Jia and Michael Backes and Neil Zhenqiang Gong and **Yang Zhang**. Stealing Links from Graph Neural Networks. In *USENIX Security Symposium (USENIX Security)*, pages 2669–2686. USENIX, 2021.

[120] Zhikun Zhang and Tianhao Wang and Jean Honorio and Ninghui Li and Michael Backes and Shibo He and Jiming Chen and **Yang Zhang**. PrivSyn: Differentially Private Data Synthesis. In *USENIX Security Symposium (USENIX Security)*, pages 929–946. USENIX, 2021.

[121] Fatemeh Tahmasbi and Leonard Schild and Chen Ling and Jeremy Blackburn and Gianluca Stringhini and **Yang Zhang** and Savvas Zannettou. "Go eat a bat, Chang!": On the Emergence of Sinophobic Behavior on Web Communities in the Face of COVID-19. In *The Web Conference (WWW)*. ACM, 2021.

[122] Rui Wen and Yu Yu and Xiang Xie and **Yang Zhang**. LEAF: A Faster Secure Search Algorithm via Localization, Extraction, and Reconstruction. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1219–1232. ACM, 2020.

[123] Dingfan Chen and Ning Yu and **Yang Zhang** and Mario Fritz. GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 343–362. ACM, 2020.

[124] Ahmed Salem and Apratim Bhattacharya and Michael Backes and Mario Fritz and **Yang Zhang**. Updates-Leak: Data Set Inference and Reconstruction Attacks in Online Learning. In *USENIX Security Symposium (USENIX Security)*, pages 1291–1308. USENIX, 2020.

[125] Inken Hagestedt and Mathias Humbert and Pascal Berrang and Irina Lehmann and Roland Eils and Michael Backes and **Yang Zhang**. Membership Inference Against DNA Methylation Databases. In *IEEE European Symposium on Security and Privacy (Euro S&P)*, pages 509–520. IEEE, 2020.

[126] **Yang Zhang** and Mathias Humbert and Bartlomiej Surma and Praveen Manoharan and Jilles Vreeken and Michael Backes. Towards Plausible Graph Anonymization. In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2020.

[127] Jinyuan Jia and Ahmed Salem and Michael Backes and **Yang Zhang** and Neil Zhenqiang Gong. MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 259–274. ACM, 2019.

[128] Zheng Li and Chengyu Hu and **Yang Zhang** and Shanqing Guo. How to Prove Your Model Belongs to You: A Blind-Watermark based Framework to Protect Intellectual Property of DNN. In *Annual Computer Security Applications Conference (ACSAC)*, pages 126–137. ACSAC, 2019.

[129] Zhiqiang Zhong and **Yang Zhang** and Jun Pang. A Graph-Based Approach to Explore Relationship Between Hashtags and Images. In *International Conference Web Information Systems Engineering (WISE)*, pages 473–488. Springer, 2019.

[130] Tahleen Rahman and Bartlomiej Surma and Michael Backes and **Yang Zhang**. Fairwalk: Towards Fair Graph Embedding. In *International Joint Conferences on Artifical Intelligence (IJCAI)*, pages 3289–3295. IJCAI, 2019.

[131] **Yang Zhang**. Language in Our Time: An Empirical Analysis of Hashtags. In *The Web Conference (WWW)*, pages 2378–2389. ACM, 2019.

[132] Ahmed Salem and **Yang Zhang** and Mathias Humbert and Pascal Berrang and Mario Fritz and Michael Backes. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2019.

[133] Inken Hagestedt and **Yang Zhang** and Mathias Humbert and Pascal Berrang and Haixu Tang and XiaoFeng Wang and Michael Backes. MBeacon: Privacy-Preserving Beacons for DNA Methylation Data. In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2019.

[134] Fanghua Zhao and Linan Gao and **Yang Zhang** and Zeyu Wang and Bo Wang and Shanqing Guo. You Are Where You App: An Assessment on Location Privacy of Social Applications. In *International Symposium on Software Reliability Engineering (ISSRE)*, pages 236–247. IEEE, 2018.

[135] **Yang Zhang** and Mathias Humbert and Tahleen Rahman and Cheng-Te Li and Jun Pang and Michael Backes. Tagvisor: A Privacy Advisor for Sharing Hashtags. In *The Web Conference (WWW)*, pages 287–296. ACM, 2018.

[136] Pascal Berrang and Mathias Humbert and **Yang Zhang** and Irina Lehmann and Roland Eils and Michael Backes. Dissecting Privacy Risks in Biomedical Data. In *IEEE European Symposium on Security and Privacy (Euro S&P)*, pages 62–76. IEEE, 2018.

[137] Michael Backes and Mathias Humbert and Jun Pang and **Yang Zhang**. walk2friends: Inferring Social Links from Mobility Profiles. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1943–1957. ACM, 2017.

[138] Jun Pang and **Yang Zhang**. Quantifying Location Sociality. In *ACM Conference on Hypertext and Social Media (HT)*, pages 145–154. ACM, 2017.

[139] Jun Pang and **Yang Zhang**. DeepCity: A Feature Learning Framework for Mining Location Check-Ins. In *International Conference on Weblogs and Social Media (ICWSM)*, pages 652–655. AAAI, 2017.

[140] Yan Wang and Zongxu Qin and Jun Pang and **Yang Zhang** and Xin Jin. Semantic Annotation for Places in LBSN Using Graph Embedding. In *ACM International Conference on Information and Knowledge Management (CIKM)*, page 2343–2346. ACM, 2017.

[141] **Yang Zhang** and Minyue Ni and Weili Han and Jun Pang. Does #like4like Indeed Provoke More Likes? In *International Conference on Web Intelligence (WI)*, pages 179–186. ACM, 2017.

[142] Minyue Ni and **Yang Zhang** and Weili Han and Jun Pang. An Empirical Study on User Access Control in Online Social Networks. In *ACM Symposium on Access Control Models and Technologies (SACMAT)*, pages 12–23. ACM, 2016.

[143] Jun Pang and Polina Zablotskaia and **Yang Zhang**. On Impact of Weather on Human Mobility in Cities. In *International Conference Web Information Systems Engineering (WISE)*, pages 247–256. Springer, 2016.

[144] Jun Pang and **Yang Zhang**. Location Prediction: Communities Speak Louder than Friends. In *ACM Conference on Online Social Networks (COSN)*, pages 161–171. ACM, 2015.

[145] **Yang Zhang** and Jun Pang. Distance and Friendship: A Distance-based Model for Link Prediction in Social Networks. In *Asia-Pacific Web Conference (APWeb)*, pages 55–66. Springer, 2015.

[146] Jun Pang and **Yang Zhang**. Event Prediction with Community Leaders. In *Conference on Availability, Reliability and Security (ARES)*, pages 238–243. IEEE, 2015.

[147] Marcos Cramer and Jun Pang and **Yang Zhang**. A Logical Approach to Restricting Access in Online Social Networks. In *ACM Symposium on Access Control Models and Technologies (SACMAT)*, pages 75–86. ACM, 2015.

[148] Jun Pang and **Yang Zhang**. Cryptographic Protocols for Enforcing Relationship-based Access Control Policies. In *Annual IEEE Computers, Software and Applications Conference (COMPSAC)*, pages 484–493. IEEE, 2015.

[149] Jun Pang and **Yang Zhang**. Exploring Communities for Effective Location Prediction. In *International Conference on World Wide Web (WWW)*, pages 87–88. ACM, 2015.

[150] **Yang Zhang** and Jun Pang. Community-driven Social Influence Analysis and Applications. In *International Conference on Web Engineering (ICWE)*. Springer, 2015.

[151] Jun Pang and **Yang Zhang**. A New Access Control Scheme for Facebook-style Social Networks. In *Conference on Availability, Reliability and Security (ARES)*, pages 1–10. IEEE, 2014.

Journal

[152] Ziqing Yang and Rui Wen and Xinlei He and Yun Shen and Michael Backes and **Yang Zhang**. BadBone: Backdoor Attacks Against Backbone Models in Visual Prompt Learning. *IEEE Transactions on Information Forensics and Security*, 2026.

[153] Rui Zhang and Yun Shen and Hongwei Li and Wenbo Jiang and Hanxiao Chen and Yuan Zhang and Guowen Xu and **Yang Zhang**. Backdoor Complications: A Comprehensive Analysis and Mitigation of the Unforeseen Consequences of Backdoor Attacks. *IEEE Transactions on Dependable and Secure Computing*, 2026.

[154] Zheng Li and Siyuan Wu and Ruichuan Chen and Paarijaat Aditya and Istemi Ekin Akkus and Manohar Vanga and Min Zhang and Hao Li and **Yang Zhang**. Model Hijacking Attack in Federated Learning. *IEEE Transactions on Information Forensics and Security*, 2026.

[155] Linkang Du and Zhikun Zhang and Min Chen and Mingyang Sun and Shouling Ji and Peng Cheng and Jiming Chen and Michael Backes and **Yang Zhang**. Revealing the Risk of Hyper-parameter Leakage in Deep Reinforcement Learning Models. *IEEE Transactions on Dependable and Secure Computing*, 2026.

[156] Yugeng Liu and Tianshuo Cong and Zhengyu Zhao and Michael Backes and Yun Shen and **Yang Zhang**. Robustness Over Time: Understanding Adversarial Examples' Effectiveness on Longitudinal Versions of Large Language Models. *IEEE Transactions on Information Forensics and Security*, 2026.

[157] Yan Pang and Tianhao Wang and Xuhui Kang and Mengdi Huai and **Yang Zhang**. White-box Membership Inference Attacks against Diffusion Models. *Proceedings on Privacy Enhancing Technologies*, 2025.

[158] Joann Qiongna Chen and Xinlei He and Zheng Li and **Yang Zhang** and Zhou Li. A Comprehensive Study of Privacy Risks in Curriculum Learning. *Proceedings on Privacy Enhancing Technologies*, 2025.

[159] Xiaokuan Zhang and **Yang Zhang** and Yinqian Zhang. VERITRAIN: Validating MLaaS Training Efforts via Anomaly Detection. *IEEE Transactions on Dependable and Secure Computing*, 2024.

[160] Yixin Wu and Xinlei He and Pascal Berrang and Mathias Humbert and Michael Backes and Neil Zhenqiang Gong and **Yang Zhang**. Link Stealing Attacks Against Inductive Graph Neural Networks. *Proceedings on Privacy Enhancing Technologies*, 2024.

[161] Cheng-Te Li and Cheng Hsu and **Yang Zhang**. FairSR: Fairness-aware Sequential Recommendation through Multi-Task Learning with Preference Graph Embeddings. *ACM Transactions on Intelligent Systems and Technology*, 2022.

[162] Xinlei He and Qingyuan Gong and Yang Chen and **Yang Zhang** and Xin Wang and Xiaoming Fu. DatingSec: Detecting Malicious Accounts in Dating Apps Using a Content-Based Attention Network. *IEEE Transactions on Dependable and Secure Computing*, 2021.

[163] Bo-Heng Chen and Cheng-Te Li and Kun-Ta Chuang and Jun Pang and **Yang Zhang**. An Active Learning-based Approach for Location-aware Acquaintance Inference. *Knowledge and Information Systems*, 2018.

[164] Jun Pang and **Yang Zhang**. A New Access Control Scheme for Facebook-style Social Networks. *Computers & Security*, 2015.

TrustAIRLab

Our lab is fully committed to open science, which led to the establishment of TrustAIRLab.

- Much of the code developed by our lab is accessible through our GitHub organization
- A curated selection of datasets collected by our lab can be found on our Hugging Face organization and our Zenodo community

Teaching

- 2026 Summer, Seminar: Data-driven Understanding of the Disinformation Epidemic
- 2025 Winter, Seminar: AI Safety
- 2025 Summer, Advanced Lecture: Attacks Against Machine Learning Models
- 2025 Summer, Seminar: Data-driven Understanding of the Disinformation Epidemic
- 2024 Winter, Seminar: Privacy of Machine Learning
- 2024 Summer, Advanced Lecture: Attacks Against Machine Learning Models
- 2024 Summer, Seminar: Data-driven Understanding of the Disinformation Epidemic
- 2023 Winter, Seminar: Privacy of Machine Learning
- 2023 Summer, Advanced Lecture: Attacks Against Machine Learning Models
- 2023 Summer, Seminar: Data-driven Understanding of the Disinformation Epidemic
- 2022 Winter, Seminar: Privacy of Machine Learning
- 2022 Summer, Advanced Lecture: Machine Learning Privacy
- 2022 Summer, Seminar: Data-driven Understanding of the Disinformation Epidemic
- 2021 Winter, Seminar: Privacy of Machine Learning
- 2021 Summer, Advanced Lecture: Privacy Enhancing Technologies
- 2021 Summer, Seminar: Data-driven Understanding of the Disinformation Epidemic
- 2020 Winter, Seminar: Data Privacy
- 2020 Summer, Advanced Lecture: Privacy Enhancing Technologies
- 2020 Summer, Seminar: Data-driven Approaches on Understanding Disinformation
- 2019 Winter, Seminar: Data Privacy
- 2019 Summer, Advanced Lecture: Privacy Enhancing Technologies
- 2019 Summer, Seminar: Biomedical Privacy
- 2018 Winter, Seminar: Data Privacy
- 2018 Summer, Advanced Lecture: Privacy Enhancing Technologies
- 2018 Summer, Seminar: Adversarial Machine Learning

Students

Postdoc

Zihan Wang	**11/2025 -**
Mingjie Li	**10/2023 -**

Ph.D. Students

Xinyu Zhang	**6/2026 -**
Shengyun Si	**3/2026 -**
Yage Zhang	**8/2025 -**
Bo Shao	**5/2025 -**
Mengfei Liang	**5/2025 -**
Zeyuan Chen	**3/2025 -**
Ye Leng	**10/2024 -**
Yukun Jiang	**6/2024 -**
Chi Cui	**4/2024 -**
Yicong Tan	**4/2024 -**
Junjie Chu	**11/2022 -**
Ziqing Yang	**11/2022 -**
Yugeng Liu	**1/2022 -**
Boyang Zhang	**12/2021 -**
Hai Huang	**11/2021 -**
Yihan Ma	**7/2021 -**

Ph.D. Preparatory Phase

Tianze Chang **5/2025 -**

Alumni

Xinyue Shen *Ph.D. student* **10/2022 - 9/2026**
now assistant professor at University of Waterloo
(Abbe Grant 2024, KAUST Rising Star in AI 2025,
MLCommons ML and Systems Rising Star 2025)

Yixin Wu *Ph.D. student* **11/2022 - 5/2026**
now assistant professor at HKUST (GZ)
(MLCommons ML and Systems Rising Star 2025,
Abbe Grant 2025, Rising Stars in EECS 2025)

Wai Man Si *Ph.D. student* **11/2021 - 5/2026**
now applied research scientist at Sands Macau

Yiting Qu *Ph.D. student* **11/2021 - 10/2025**

Chia-Yi Hsu *visiting Ph.D. student from National Yang Ming Chiao Tung University* **12/2024 - 9/2025**

Rui Wen *Ph.D. student* **10/2021 - 4/2025**
now assistant professor at Institute of Science Tokyo

Yuke Hu *visiting Ph.D. student from Zhejiang University* **5/2024 - 12/2024**
now postdoc at KAUST

Zeyang Sha *Ph.D. student* **3/2023 - 11/2024**
now senior algorithmic engineer at Ant Financial (Ant Star)

Zhuoran Liu *visiting Ph.D. student from Radboud University* **1/2024 - 4/2024**
now assistant professor at University of Amsterdam

Zheng Li *Ph.D. student* **2/2021 - 10/2023**
now full professor at Shandong University
(ERCIM WG STM Best Ph.D. Thesis Award 2024)

Xinlei He *Ph.D. student* **2/2020 - 9/2023**
now full professor at Wuhan University
(Norton Labs Graduate Fellowship 2022)

Zhengyu Zhao *postdoc* **1/2022 - 8/2023**
now full professor at Xi'an Jiaotong University

Tianshuo Cong *visiting Ph.D. student from Tsinghua University* **8/2021 - 12/2022**
now research-track professor at Shandong University

Shannon Pierson *research intern* **3/2022 - 6/2022**
now senior fellow at UC Berkeley Center for Long-Term Cybersecurity

Ahmed Salem *Ph.D. student* **2/2017 - 1/2022**
now senior researcher at Microsoft Security Response Center

Bartlomiej Surma *Ph.D. student* **10/2016 - 9/2021**
now software engineer at Google

Service

- PC Member
  - 2027: ICML (Area Chair), ICLR (Area Chair), KDD (Area Chair), AAAI (Area Chair), DAC, SaTML
  - 2026: ICML (Area Chair), ICLR (Area Chair), ACL ARR (Area Chair), CVPR, ECCV, KDD (Area Chair), AAAI (Senior PC), ACM AI Summit, MM, SaTML
  - 2025: USENIX Security, NDSS, ICML (Area Chair), NeurIPS (Area Chair), ICLR (Area Chair), ACL ARR (Area Chair), ICCV, WWW, KDD (Area Chair), SaTML
  - 2024: IEEE S&P, CCS, ICML, NeurIPS, ICLR, ACL ARR (Area Chair), CVPR, ECCV, WWW, KDD, ACSAC, SaTML
  - 2023: IEEE S&P, CCS, NDSS, ICML, NeurIPS, ICLR, WWW, KDD, SaTML
  - 2022: CCS, USENIX Security, NeurIPS, ICLR, WWW, KDD, AAAI, PETS, ASIACCS
  - 2021: CCS, USENIX Security, WWW, AAAI, Euro S&P, PETS, ASIACCS
  - 2020: CCS, WWW, ICWSM, RAID, PETS
  - 2019: CCS, ISMB/ECCB
- Editorial Board
  - SAE
    - IEEE Transactions on Information Forensics and Security (TIFS)
  - AE
    - Transactions on Machine Learning Research (TMLR)
    - IEEE Transactions on Dependable and Secure Computing (TDSC)
    - ACM Transactions on Privacy and Security (TOPS)
- Ph.D. Thesis Committee
  - Bokang Zhang, The Chinese University of Hong Kong (Shenzhen), 2026
  - Yiyi Chen, Aalborg University, 2026
  - Quentin Le Roux, University of Rennes, 2025
  - Salijona Dyrmishi, University of Luxembourg, 2024
  - Hailong Hu, University of Luxembourg, 2024
  - Bang Wu, Monash University, 2024
  - Sinem Sav, EPFL, 2023
  - Inken Hagestedt, Saarland University, 2021
  - Benjamin Zhao, University of New South Wales, 2021

Talks

Keynote

- 2026, GI Sicherheit
- 2025, International Conference on Knowledge Science, Engineering and Management (KSEM)
- 2025, Large Model Safety Workshop
- 2024, International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP)
- 2024, Australasian Conference on Information Security and Privacy (ACISP)
- 2024, ACNS Workshop on Security in Machine Learning and its Applications (SiMLA)
- 2023, Information Security Conference (ISC)
- 2023, The AsiaCCS Workshop on Secure and Trustworthy Deep Learning Systems
- 2023, Backdoor Attacks and Defenses in Machine Learning (BANDS)
- 2022, PAIR2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured

Data

Seminar and School

- 2025, Lecturer at Graz Security Week
- 2024, Talk at Nanyang Technological University
- 2023, Lecturer at Summer School on Privacy-Preserving Machine Learning
- 2023, Talk at EPFL
- 2022, Lecturer at Summer School on Privacy-Preserving Machine Learning
- 2022, Distinguished Lecture in ViSP (Vienna Cybersecurity and Privacy Research Center) Distinguished Lecture Series
- 2021, Vector Visitor Talk at Vector Institute
- 2021, Talk at Privacy and Security in ML Seminars
- 2021, Talk at Inria
- 2020, Talk at University College London

In the Press

- 8/2023, Tricks for making AI chatbots break rules are freely available online, *New Scientist*
- 8/2023, Wie Chatbots die eigenen Regeln vergessen, *Deutschlandfunk Nova*
- 12/2022, The internet loves ChatGPT, but there's a dark side to the tech, *Fast Company*
- 4/2020, As the coronavirus spreads, so does online racism targeting Asians, new research shows, *The Washington Post*