YI ZENG

(858)-952-2135 ◇ yizeng@vt.edu
Google Scholar ◇ Github ◇ LinkedIn ◇ Webpage

Focus and Strength

My research focus is committed to weaving a narrative where innovation in AI is intrinsically tied with safety and ethical considerations, covering adversarial robustness (e.g., jailbreak LLMs, backdoors, mitigations, etc.), robust watermarking (against GenAI-based perturbations, e.g., image modifications or text paraphrasing), and fairness.

Education

Virginia Polytechnic Institute and State University (Virginia Tech) *May. 2021 - May. 2025*
Ph.D. Student in Computer Engineering; Advisor: Ruoxi Jia

University of California - San Diego (UCSD) *Aug. 2019 - Mar. 2021*
Master of Science in Machine Learning and Data Science, Electrical and Computer Engineering

Xidian University (XDU) *Sep. 2015 - Jun. 2019*
Bachelor of Engineering in Electrical and Information Engineering ; Advisor: Huaxi Gu

Experience

Meta, Menlo Park *Summer, 2023*
Research Scientist Intern - Responsible AI

- Developed a meta-learning approach to improve model fairness. Worst group AUC improved from 0.53 to 0.72.
- Collaborated with legal/policy teams to improve Llama-2 models' safety alignment (red teaming, RLHF).

Sony AI, New York *Summer, 2022*
AI Research Intern - Privacy-Preserving Machine Learning

- Explored attacks and defenses to enhance model security when training on untrustworthy crowd-sourced data.
- Proposed certified robustness techniques for identifying backdoors and universal perturbations.

Recent Manuscripts

(1) Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!
— Safety Alignment of ChatGPT can be compromised with just 10 training examples, a cost of less than $0.20!
Xiangyu Qi*, **Yi Zeng***, Tinghao Xie*, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal and Peter Henderson
Preprint, project website, the New York Times Exclusive Coverage, 2023.

Publications (Selected)

(1) How to Sift Out a Clean Data Subset in the Presence of Data Poisoning?
Yi Zeng*, Minzhou Pan*, Himanshu Jahagirdar, Ming Jin, Lingjuan Lyu and Ruoxi Jia
USENIX Security Symposium (USENIX Security), 2023.

(2) NARCISSUS: A Practical Clean-Label Backdoor Attack with Limited Information
Yi Zeng*, Minzhou Pan*, Hoang Anh Just, Lingjuan Lyu, Meikang Qiu and Ruoxi Jia
ACM SIGSAC Conf. on Computer and Communications Security (ACM CCS) 2023.

(3) Revisiting Data-Free Knowledge Distillation with Poisoned Teachers
Junyuan Hong*, **Yi Zeng***, Shuyang Yu*, Lingjuan Lyu, Ruoxi Jia and Jiayu Zhou
International Conf. on Machine Learning (ICML), 2023.

(4) ASSET: Robust Backdoor Data Detection Across a Multiplicity of Deep Learning Paradigms
Minzhou Pan*, **Yi Zeng***, Lingjuan Lyu, Xue Lin and Ruoxi Jia
USENIX Security Symposium (USENIX Security), 2023.

(5) Turning a Curse into a Blessing: Enabling In-Distribution-Data-Free Backdoor Removal via Stabilized Model Inversion
Si Chen, **Yi Zeng**, Jiachen T. Wang, Won Park, Xun Chen, Lingjuan Lyu, Zhuoqing Mao and Ruoxi Jia
Transactions on Machine Learning Research, 2023.

(6) Alteration-free and Model-agnostic Origin Attribution of Generated Images
Zhenting Wang, Chen Chen, **Yi Zeng**, Lingjuan Lyu and Shiqing Ma
Advances in Neural Information Processing Systems (NeurIPS), 2023.

(7) Towards Robustness Certification Against Universal Perturbations
Yi Zeng*, Zhouxing Shi*, Ming Jin, Feiyang Kang , Lingjuan Lyu , Cho-Jui Hsieh and Ruoxi Jia
International Conf. on Learning Representations (ICLR), 2023.

(8) LAVA: Data Valuation without Pre-Specified Learning Algorithms
Hoang Anh Just*, Feiyang Kang*, Jiachen T. Wang, **Yi Zeng**, Myeongseob Ko, Ming Jin and Ruoxi Jia
Spotlight of International Conf. on Learning Representations (ICLR), 2023.

(9) CATER: Intellectual Property Protection on Text Generation APIs via Conditional Watermarks
Xuanli He*, Qiongkai Xu*, **Yi Zeng**, Lingjuan Lyu, Fangzhao Wu, Jiwei Li and Ruoxi Jia
Advances in Neural Information Processing Systems (NeurIPS), 2022.

(10) Adversarial Unlearning of Backdoors via Implicit Hypergradient
Yi Zeng, Si Chen, Won Park, Z. Morley Mao, Ming Jin and Ruoxi Jia
International Conf. on Learning Representations (ICLR), 2022.

(11) Adaptive Backdoor Trigger Detection in Edge-Deployed DNNs in 5G-Enabled IIoT Systems
Yi Zeng, Ruoxi Jia and Meikang Qiu
IEEE Transactions on Industrial Informatics, 2021.

(12) Rethinking the Backdoor Attacks' Triggers: A Frequency Perspective
Yi Zeng*, Won Park*, Z. Morley Mao and Ruoxi Jia
International Conf. on Computer Vision (ICCV), 2021.

(13) DeepSweep: An Framework for Mitigating DNN Backdoor Attacks using Data Augmentation
Han Qiu, **Yi Zeng**, Shangwei Guo, Tianwei Zhang, Meikang Qiu and Bhavani Thuraisingham
ACM Asia Conf. on Computer and Communications Security (AsiaCCS), 2021.

(14) Fine-tuning Is Not Enough: A Simple yet Effective Watermark Removal Attack for DNN Models
Shangwei Guo, Tianwei Zhang, Han Qiu, **Yi Zeng**, Tao Xiang and Yang Liu
International Joint Conf. on Artificial Intelligence (IJCAI), 2021.

(15) An Effective and Efficient Preprocessing-based Approach to Mitigate Advanced Adversarial Attacks
Han Qiu*, **Yi Zeng***, Qinkai Zheng, Tianwei Zhang, Meikang Qiu and Bhavani Thuraisingham
IEEE Transactions on Computers, 2020.

(16) Defending Adversarial Examples in Computer Vision based on Data Augmentation Techniques
Yi Zeng, Han Qiu, Gerard Memmi and Meikang Qiu
Best Paper of International Conf. on Algo & Archit for Parallel Processing (ICA3PP), 2020.

(17) Deep Learning Based Network Encrypted Traffic Classification and Intrusion Detection Framework
Yi Zeng, Huaxi Gu, Wenting Wei and Yantao Guo
IEEE Access, 2019.

(18) Senior2local: A Machine Learning Based Intrusion Detection Method for VANETs
Yi Zeng, Meikang Qiu, Zhong Ming and Meiqin Liu
International Conf. on Smart Computing and Communication (SmartCom), Springer, 2018.

Honors

- Amazon Ph.D. Fellowship, Amazon.com, Inc. *2022*
- Best Paper Award, 20th ICA3PP. *2020*
- Outstanding Senior Thesis Award, Xidian University. *2019*
- Outstanding Academic Scholarship, Xidian University. *2015, 2016, 2017, 2018*

Books

(1) Research and Technical Writing for Science and Engineering
Meikang Qiu, Han Qiu and **Yi Zeng**
CRC Press, 2022.

Proposals and Grants

(1) Annual Competition on Emerging Issues of Data Security and Privacy
Yi Zeng, Meikang Qiu and Ruoxi Jia
Grants for Emerging Technology Activities, IEEE Computer Society, 2022.

Academic Services

Competition Chair:	IEEE Trojan Removal Competition (reports: PR Newswire), 2022
Conf. Reviewer/PC:	CVPR-24,23,22 (outstanding), ICLR-24, NeurIPS-23,22, ICML-23,22, ICCV-23, ECCV-22, AAAI-22, KSEM-22, KSEM-21, EUC-21, IEEE ISPA-21, ICA3PP-20
Journal Reviewer :	TPAMI, IEEE TNNLS, IEEE TDSC, IEEE TII, VEHCOM