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 IP watermarking (against GenAI-based perturbations, e.g., image modifications or text paraphrasing), trustworthiness, and fairness.

Education

Virginia Polytechnic Institute and State University (Virginia Tech) — *May. 2021 - May. 2025*
Doctor of Philosophy 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

Industry Experience

Virtue AI, San Francisco - Founding Research Scientist - Red-Teaming & Mitigation — *Summer 2024 - Present*

- Spearheaded the development of one of the most comprehensive, regulation-aligned AI safety risk taxonomy and benchmarking evaluation pipeline, encompassing **320+** unique risk categories derived from regulations, company policies, and tailored use cases (e.g., hallucination, brand risk).
- Led the creation of *VirtueRed*, an innovative automated and adaptive red-teaming agent platform leveraging quality diversity algorithms, resulting in a **5.75**× increase in red-teaming success rate for `Llama-3.1-8b-Instruct` compared to the GCG attack.
- Orchestrated a team of five in developing a cutting-edge suite of datasets and tools for red-teaming and capability evaluation of audio-based foundation models (delivered to **OpenAI**), focusing on key aspects such as voice recognition, multi-participant identification, adversarial testing, and safety jailbreaking.

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

- Pioneered a novel training-time meta-learning approach to enhance model fairness, addressing multiple fairness criteria including group parity, equal presenting, and Rawlsian Max-min fairness. Achieved a significant improvement in worst-group AUC from **0.53 to 0.72** while maintaining overall model performance.
- Collaborated cross-functionally with legal and policy teams to bolster `Llama-2` models' safety alignment, conducting comprehensive red-teaming studies utilizing gradient-based methods and social science-based techniques.

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

- Advanced novel attack methodologies in-ward pointing gradient searching, developing clean-label data poisoning attacks that surpassed previous state-of-the-art methods by a factor of **50**× in success rate.
- Engineered robust defense mechanisms, including a generalizable data sifting approach achieving **100%** clean selection in 100-size sets, and training-time defenses for various paradigms (supervised, self-supervised, transfer learning), reducing attack success rates from over 90% to below **2%** on untrustworthy crowd-sourced data.
- Developed a certified robustness pipeline providing up to **23.32%** tighter certified robust rates against backdoors and universal perturbations compared to sample-wise certifications, significantly enhancing model security guarantees.

Publications (Selected)

(1) BEEAR: Embedding-based Adversarial Removal of Safety Backdoors in Instruction-tuned Language Models
Yi Zeng*, Weiyu Sun*, Tran Ngoc Huynh, Dawn Song, Bo Li and Ruoxi Jia
EMNLP, 2024.

(2) How Johnny Can Persuade LLMs to Jailbreak Them: Rethinking Persuasion to Challenge AI Safety by Humanizing LLMs
Yi Zeng*, Hongpeng Lin*, Jingwen Zhang, Diyi Yang, Ruoxi Jia and Weiyan Shi
Best Social Impact Paper Award at ACL, 2024.

(3) RigorLLM: Resilient Guardrails for Large Language Models against Undesired Content
Zhuowen Yuan, Zidi Xiong, **Yi Zeng**, Ning Yu, Ruoxi Jia, Dawn Song and Bo Li
ICML, 2024.

(4) A Safe Harbor for AI Evaluation and Red Teaming
Shayne Longpre, Sayash Kapoor, Kevin Klyman, Ashwin Ramaswami, Rishi Bommasani, Borhane Blili-Hamelin, Yangsibo Huang, Aviya Skowron, Zheng Xin Yong, Suhas Kotha, **Yi Zeng**, Weiyan Shi, Xianjun Yang, Reid Southen, Alexander Robey, Patrick Chao, Diyi Yang, Ruoxi Jia, Daniel Kang, Alex Pentland, Arvind Narayanan, Percy Liang and Peter Henderson
Oral presentation at ICML, 2024.

(5) Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!
Xiangyu QI*, **Yi Zeng***, Tinghao Xie*, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal and Peter Henderson
Oral presentation at ICLR, 2024. Featured in *New York Times*. Highlighted in *NIST AI 800-1*.

(6) 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, 2023.

(7) 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 CCS, 2023.

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

(9) 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, 2023.

(10) 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.

(11) Alteration-free and Model-agnostic Origin Attribution of Generated Images
Zhenting Wang, Chen Chen, **Yi Zeng**, Lingjuan Lyu and Shiqing Ma
NeurIPS, 2023.

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

(13) 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 presentation at ICLR, 2023.

(14) 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
NeurIPS, 2022.

(15) Adversarial Unlearning of Backdoors via Implicit Hypergradient
Yi Zeng, Si Chen, Won Park, Z. Morley Mao, Ming Jin and Ruoxi Jia
ICLR, 2022. Highlighted in *NIST AI 100-2e2023*.

(16) Rethinking the Backdoor Attacks' Triggers: A Frequency Perspective
Yi Zeng*, Won Park*, Z. Morley Mao and Ruoxi Jia
ICCV, 2021.

(17) DeepSweep: An Framework for Mitigating DNN Backdoor Attacks using Data Augmentation
Han Qiu, **Yi Zeng**, Shangwei Guo, Tianwei Zhang, Meikang Qiu and Bhavani Thuraisingham
AsiaCCS, 2021.

(18) 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.

(19) Defending Adversarial Examples in Computer Vision based on Data Augmentation Techniques
Yi Zeng, Han Qiu, Gerard Memmi and Meikang Qiu
Best Paper Award at ICA3PP, 2020.

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

Recent Manuscripts (selected)

(1) WokeyTalky: Towards Scalable Evaluation of Misguided Safety Refusal in LLMs
Yi Zeng*, Adam Nguyen, Bo Li and Ruoxi Jia
Preprint, Project page, GitHub, Dataset, 2024.

(2) RedCode: Risky Code Execution and Generation Benchmark for Code Agents
Chengquan Guo*, Xun Liu*, Chulin Xie*, Andy Zhou, **Yi Zeng**, Zinan Lin, Dawn Song and Bo Li
Preprint, Leaderboard, 2024.

(3) AIR-Bench 2024: A Safety Benchmark Based on Risk Categories from Regulations and Policies
Yi Zeng*, Yu Yang*, Andy Zhou*, Jeffrey Ziwei Tan*, Yuheng Tu*, Yifan Mai*, Kevin Klyman, Minzhou Pan, Ruoxi Jia, Dawn Song, Percy Liang, Bo Li
Preprint, Leaderboard, Dataset, Featured in *Wired* , 2024.

(4) AI Risk Categorization Decoded (AIR 2024): From Government Regulations to Corporate Policies
Yi Zeng*, Kevin Klyman*, Andy Zhou, Yu Yang, Minzhou Pan, Ruoxi Jia, Dawn Song, Percy Liang and Bo Li
Preprint, Blog Post, Featured in *Wired* , 2024.

(5) Sorry-bench: Systematically evaluating large language model safety refusal behaviors
Tinghao Xie*, Xiangyu Qi*, **Yi Zeng***, Yangsibo Huang*, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson and Prateek Mittal
Preprint, Project page, Dataset, 2024.

(6) Fairness-Aware Meta-Learning via Nash Bargaining
Yi Zeng*, Xuelin Yang*, Li Chen, Cristian Canton Ferrer, Ming Jin, Michael I Jordan and Ruoxi Jia
Preprint, GitHub, 2024.

(7) JIGMARK: A Black-Box Approach for Enhancing Image Watermarks against Diffusion Model Edits
Minzhou Pan*, **Yi Zeng***, Ning Yu, Cho-Jui Hsieh, Peter Henderson, Ruoxi Jia and Xue Lin
Preprint, GitHub, 2024.

Honors

- Best Social Impact Paper Award, 62th ACL. *2024*
- 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*

Invited Talks, Lectures & Panels

- "AI Safety: Are the Guardrails Built Around A.I. Systems Sufficient?" Atlas of AI Summit, Capital One Hall. *2024*
- "Unboxing AI Risks in the Age of LLMs" AI Expo for National Competitiveness, Washington Convention Center. *2024*
- "Decoding AI Safety: Navigating the Multifaceted Dimensions towards LLM safety" Guest Lecture, University of Chicago. *2024*
- "LLM Safety: Current Status and Future Outlook" Bank of New York Mellon, AI Hub team. *2024*
- "Rethinking Persuasion to Challenge AI Safety by Humanizing LLMs" Meta Platforms, Inc., GenAI safety team. *2024*
- "Data-centric Backdoor Attacks and Countermeasures" TMLR Young Scientist Seminar, Hong Kong Baptist University. *2023*
- "The Duel of Spear and Shield—Neural Network Backdoor Attack and Defense" AI TIME Ph.D. Debate Panel, online. *2022*

Proposals and Grants

(1) Adaptive Safety Framework for Secure and Responsible CodeGen LLMs
Yi Zeng and Ruoxi Jia
$250,000 , Amazon Trusted AI Challenge, Amazon Science, 2024.

(2) Annual Competition on Emerging Issues of Data Security and Privacy
Yi Zeng, Meikang Qiu and Ruoxi Jia
$50,000 , Grants for Emerging Technology Activities, IEEE Computer Society, 2022.

Academic Services

Competition Chair: The Competition for LLM and Agent Safety 2024 (NeurIPS CLAS 2024), 2024
IEEE Trojan Removal Competition (IEEE TRC'22) (reports: PR Newswire), 2022
Workshop Chair: Trustworthy Interactive Decision-Making with Foundation Models Workshop (IJCAI TIDMwFM), 2024
Conf. Reviewer/PC: CVPR-24,23,22 (outstanding), ICLR-24, NeurIPS-24,23,22, ICML-24,23,22, 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