Ren Pang

(484)747–2401 | ain-soph@live.com | ain-soph.github.io

EDUCATION

Ph.D.	*Information Sciences and Technology*	Pennsylvania State University	2019 – 2023
B.Sc.	*Mathematics*	Nankai University	2014 – 2018

WORK EXPERIENCE

Applied Scientist, *Amazon* 2024 – Present

Develop Guardrails for Amazon Bedrock, the safeguards customized to customers' application requirements and responsible AI policies.

INTERN EXPERIENCE

Applied Scientist (Intern), *Amazon* 2023 Summer

Explore the vulnerabilities of LLMs to jail-breaking attacks, where Reinforcement Learning from Human Feedback (RLHF) is considered to enhance the attack and defense efficiency. During project development, I submitted several bug fixes and new features to Transformers, Peft and Trl libraries.

Machine Learning Engineer (Intern), *Meta* 2022 Summer

Pages and Groups Integrity: Introduce new classification model for malicious page detection. It mitigates the impact of incorrect label annotation, and provides interpretable classification outputs for better user experience.

TorchVision: Provide the official TorchVision implementation of SwinTransformerV2.

PUBLICATIONS

1. On the Difficulty of Defending Contrastive Learning against Backdoor Attacks,
   C. Li, **R. Pang**, B. Cao, J. Chen, S. Ji, and T. Wang,
   Proceedings of *the USENIX Security Symposium* (**USENIX**), 2024.
2. Defending Pre-trained Language Models as Few-shot Learners Against Backdoor Attacks,
   Z. Xi, T. Du, C. Li, **R. Pang**, S. Ji, J. Chen, F. Ma, and T. Wang,
   Proceedings of *Advances in Neural Information Processing Systems* (**NeurIPS**), 2023.
3. An Embarrassingly Simple Backdoor Attack against Self-supervised Learning,
   C. Li, **R. Pang**, Z. Xi, T. Du, S. Ji, Y. Yao, and T. Wang,
   Proceedings of *the International Conference on Computer Vision* (**ICCV**), 2023.
4. On the Security Risks of Knowledge Graph Reasoning,
   Z. Xi, T. Du, C. Li, **R. Pang**, S. Ji, X. Luo, X. Xiao, F. Ma, and T. Wang,
   Proceedings of *the USENIX Security Symposium* (**USENIX**), 2023.
5. The Dark Side of AutoML: Towards Architectural Backdoor Search,
   R. Pang, C. Li, Z. Xi, S. Ji, and T. Wang,
   Proceedings of *the International Conference on Learning Representations* (**ICLR**), 2023.
6. TrojanZoo: Towards Unified, Holistic, and Practical Evaluation of Neural Backdoors,
   R. Pang, Z. Zhang, X. Gao, Z. Xi, S. Ji, P. Cheng, and T. Wang,
   Proceedings of *the IEEE European Symposium on Security and Privacy* (**EuroS&P**), 2022.
7. On the Security Risks of AutoML,
   R. Pang, Z. Xi, S. Ji, X. Luo, and T. Wang,
   Proceedings of *the USENIX Security Symposium* (**USENIX**), 2022.
8. Graph Backdoor,
   Z. Xi, **R. Pang**, S. Ji, and T. Wang,
   Proceedings of *the USENIX Security Symposium* (**USENIX**), 2021.

9. i-Algebra: Towards Interactive Interpretability of Deep Neural Networks,
X. Zhang, **R. Pang**, S. Ji, F. Ma, and T. Wang,
Proceedings of *the AAAI Conference on Artificial Intelligence* (**AAAI**), 2021.
10. AdvMind: Inferring Adversary Intent of Black-Box Attacks,
R. Pang, X. Zhang, S. Ji, X. Luo, and T. Wang,
Proceedings of *the ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (**KDD**), 2020.
11. A Tale of Evil Twins: Adversarial Inputs versus Poisoned Models,
R. Pang, H. Shen, X. Zhang, S. Ji, Y. Vorobeychik, X. Luo, A. Liu, and T. Wang,
Proceedings of *the ACM Conference on Computer and Communications Security* (**CCS**), 2020.

OPEN-SOURCE CONTRIBUTION

1. **TrojanZoo** (*owner*) https://github.com/ain-soph/trojanzoo
   (70,000 Lines) Offer a universal, flexible PyTorch platform to conduct security analysis of attacks and defenses on deep neural network models.
2. **AlpsPlot** (*owner*) https://github.com/ain-soph/alpsplot
   (14,000 Lines) Offer a high-level python library to plot academic figures based on Matplotlib.
3. **TorchVision.SwinTransformerV2** https://github.com/pytorch/vision/pull/6246
   (400 Lines) Provide TorchVision official implementation of SwinTransformerV2.
4. **Matplotlib.Text** https://github.com/matplotlib/matplotlib/pull/20101
   Fix Text class bug when font argument is provided without math_fontfamily.