Shengfang Zhai (翟胜方)

Phone: (+86) 17801180535 / (+65) 82155209 | Email: zhaisf@stu.pku.edu.cn, shengfang.zhai@gmail.com

Address: Peking University, Beijing, China, 100871 | Google scholar | Homepage (https://zhaisf.github.io/)

Personal Summary

I am currently a PhD candidate at Peking University. My research interests primarily focus on the security and privacy issues associated with generative models, particularly diffusion models and large language models (LLMs).

I am highly motivated and willing to learning new things, resistant to pressure. In personal life, I am healthy and athletic, and enjoy running and playing basketball.

Research interest: AI Security & Privacy, Generative Models, Diffusion Models, LLM

Education and Working Experience

Research intern in National University of Singapore Singapore
Advisor: Prof. Jiaheng Zhang 2024.8 – Present

- Work on the security/privacy/copyright issues of text-to-image diffusion models, large language models.

Visiting Ph.D. student in Nanyang Technological University Singapore
Advisor: Prof. Yang Liu 2023.12 – 2024.8

- Work on the privacy issues of text-to-image diffusion models (*NeurIPS'24*).

Research intern in Tsinghua University (*TSAIL*) Beijing, China
Advisor: Prof. Hang Su and Dr. Yinpeng Dong 2022.9 – 2023.12

- Investigate the backdoor threats against text-to-image diffusion models (*ACM MM'23*).

Peking University, Ph.D. in Software Engineering (Recommended) Beijing, China
Advisor: Prof. Qingni Shen 2020.9 - Present

- Honors and Awards: Merit Student, Academic Excellence Award (Top 1%), Shenzhen Stock Exchange Scholarship (Top 5%).

China Agricultural University, B.S. in Computer Science and Technology ***(Honored Program)*** Beijing, China
2016.9 – 2020.6

- Honors and Awards: Merit Student, Academic Excellence Award, Science Base Class Scholarship (3%), First prize in mathematics competition.

Selected Papers

Publications (Conferences)

1. Membership Inference on Text-to-Image Diffusion Models via Conditional Likelihood Discrepancy [URL]

   Shengfang Zhai, Huanran Chen, Yinpeng Dong, Jiajun Li, Qingni Shen, Yansong Gao, Hang Su, Yang Liu

   Advances in Neural Information Processing Systems **(NeurIPS, CCF-A),** 2024

   (**TL; DR:** We propose the membership inference on text-to-image diffusion models via condition likelihood discrepancy, outperforming previous works on diverse datasets, with superior resistance against early stopping and data augmentation.)

2. Text-to-image diffusion models can be easily backdoored through multimodal data poisoning [URL]

   Shengfang Zhai, Yinpeng Dong, Qingni Shen, Shi Pu, Yuejian Fang, Hang Su.

   ACM International Conference on Multimedia (**ACM MM, Oral, CCF-A**), 2023

   (**TL; DR:** We pioneer the investigation of backdoor attack techniques on text-to-image diffusion models.)

3. NCL: Textual Backdoor Defense Using Noise-augmented Contrastive Learning [URL]

   Shengfang Zhai, Qingni Shen, Xiaoyi Chen, Weilong Wang, Cong Li, Yuejian Fang, Zhonghai Wu.

IEEE International Conference on Acoustics, Speech, and Signal Processing (**ICASSP, CCF-B**), 2023

4. Kallima: A Clean-label Framework for Textual Backdoor Attacks [URL]
   Xiaoyi Chen, Yinpeng Dong, Zeyu Sun, **Shengfang Zhai**, Qingni Shen, and Zhonghai Wu.
   European Symposium on Research in Computer Security (**ESORICS, CCF-B**), 2022
5. Automated extraction of abac policies from natural-language documents in healthcare systems [URL]
   Yutang Xia, **Shengfang Zhai**, Qinting Wang, Huiting Hou, Zhonghai Wu, Qingni Shen.
   IEEE International Conference on Bioinformatics and Biomedicine (**BIBM, CCF-B**), 2022

Selected Program Works

LLM Unlearning (Student Leader) Peking University, 2024.05

We propose a machine unlearning method tailored for LLMs, capable of erasing privacy data from LLMs while preserving their utility, and preventing the meaningless tokens loop of previous LLM unlearning methods.

NLP-based cloud security standards compliance evaluation strategy Peking University, 2020.9-2021.9

For compliance issues when deploying or migrating across cloud platforms, we design NLP method to help users quickly determine whether the security standards between different cloud platforms are equivalent.

Challenges

ByteDance Security AI Challenge: Top 2% (Textual Adversarial Attack Track) 2022.10

Computer Skills

Languages: C, C++, Python, MATLAB
Deep Learning Tools: Pytorch, Tensorflow
Operating Systems: Windows, Linux + Shell, Mac OSX.

Services

Committee Members	CCS AEC 2024
Reviewer for Journals	IEEE TPAMI, IEEE TNNLS, Elsevier Computer & Security, Neurocomputing
Reviewer for Conferences	ICLR, CVPR, ACL, EMNLP, ACM MM, AAAI, AsiaCCS, ECAI, ICASSP, ICICS