Lu Yan

Location: West Lafayette, IN *E-mail:* yan390@purdue.edu *Tel:* 217-991-0003 *Website:* lunaryan.github.io

Education

Ph.D. in Computer Science *Purdue University*
West Lafayette, IN *08/2021 - 05/2026 (expected)*
Advisor: Prof. Xiangyu Zhang
Research Focus: AI Safety, specializing in LLMs and Diffusion Models

B.E. in Computer Science and Engineering *Shanghai Jiao Tong University*
Shanghai, China *09/2016 - 06/2020*
Zhiyuan Honors Program of Engineering (top 5% in SJTU)

Technical Skills

Programming Languages: Python, C/C++, Java, Bash

Frameworks & Libraries: PyTorch, QLoRA, torchtune, FSDP, DeepSpeed, TensorFlow, Transformers

Machine Learning Skills: LLMs: fine-tuning, jailbreaking, and detection; Diffusion models: training, Dreambooth, Stable Diffusion, Latent Diffusion, guided diffusion; Multi-modal models, adversarial training, backdoor detection

System Security Skills: Linux attack fundamentals, memory exploitation using pwntools, automated program analysis, fuzzing, symbolic execution

Publications

▸PREPRINT **ASPIRER: Bypassing System Prompts with Permutation-based Backdoors in LLMs**

Lu Yan, Siyuan Cheng, Xuan Chen, Kaiyuan Zhang, Guangyu Shen, Zhuo Zhang, Xiangyu Zhang

TL;DR: We introduce the first work on systematically bypassing system prompts in LLMs and propose permutation triggers, which activate only when specific components are ordered correctly. Our attack is stealthy and adaptive to unforeseen user prompts. ASPIRER achieves up to 100% ASR and CACC, demonstrating robust performance across diverse scenarios.

▸PREPRINT **RL-JACK: Reinforcement Learning-powered Black-box Jailbreaking Attack against LLMs**

Xuan Chen, Yuzhou Nie, **Lu Yan**, Yunshu Mao, Wenbo Guo, Xiangyu Zhang

TL;DR: RL-JACK is the first DRL-driven black-box jailbreaking attack against LLMs. It employs an agent to select strategies while a helper LLM generates prompts, reducing action space and enabling consistent strategy learning. This approach achieves up to +30% ASR over SOTA attacks and transfers across LLMs with 95-97% effectiveness.

▸IEEE S&P (OAKLAND) 2025 **BAIT: Large Language Model Backdoor Scanning by Inverting Attack Target**

Guangyu Shen, Siyuan Cheng, Zhuo Zhang, Guanhong Tao, Kaiyuan Zhang, Hanxi Guo, **Lu Yan**, Xiaolong Jin, Shengwei An, Shiqing Ma, Xiangyu Zhang

TL;DR: This work proves strong correlations between target tokens in backdoored LLM outputs, even without triggers. Based on this, BAIT scans backdoors in LLMs by inverting target outputs and achieves 0.98 ROC-AUC across 125 models with only black-box access, significantly outperforming five state-of-the-art baselines.

▸USENIX SECURITY 2024 **Rethinking the invisible protection against unauthorized image usage in stable diffusion**

Shengwei An*, **Lu Yan***, Siyuan Cheng, Guangyu Shen, Kaiyuan Zhang, Qiuling Xu, Guanhong Tao, Xiangyu Zhang (*equal contribution)

TL;DR: INSIGHT defeats protections that add invisible noise to images by leveraging the fact that these protections, invisible to human eyes, also lose effectiveness in photos. INSIGHT aligns the protected image with a reference photo in both VAE and UNet stages, demonstrating 1.4x effectiveness and outperforming four SOTA baselines in 93.9% cases in a user study.

▸NEURIPS 2023 **ParaFuzz: An Interpretability-Driven Technique for Detecting Poisoned Samples in NLP**

Lu Yan, Zhuo Zhang, Guanhong Tao, Kaiyuan Zhang, Xuan Chen, Guangyu Shen, Xiangyu Zhang

TL;DR: Leading the way in applying fuzzing to NLP security, PARAFUZZ detects poisoned samples by using ChatGPT to paraphrase and remove triggers. It optimizes paraphrasing prompts via fuzzing, introducing sentence coverage and three mutation strategies. PARAFUZZ achieves a 90.1% F1 score, more than double that of most baselines.

▸PREPRINT **Rapid optimization for jailbreaking llms via subconscious exploitation and echopraxia**

Guangyu Shen, Siyuan Cheng, Kaiyuan Zhang, Guanhong Tao, Shengwei An, **Lu Yan**, Zhuo Zhang. Shiqing Ma, Xiangyu Zhang

TL;DR: Ripple introduces a novel jailbreaking approach, extracting hidden malicious knowledge suppressed by alignment protections by interrogation. Ripple achieves a +42.18% higher success rate compared to state-of-the-art baselines, and exhibits strong transferability across both open-source and commercial LLMs.

▸BUGS @ NEURIPS 2023 **D^3: Detoxing Deep Learning Dataset**

Lu Yan, Siyuan Cheng, Guangyu Shen, Guanhong Tao, Xuan Chen, Kaiyuan Zhang, Yunshu Mao, Xiangyu Zhang

TL;DR: D^3 is a dataset detoxification tool that first extracts poison triggers using differential analysis and dual-tanh perturbations. After extraction, it trains a classifier that identifies poisoned data. D^3 achieves over 95% precision and 95% recall across 42 poisoned datasets, outperforming state-of-the-art methods.

▸ESEC/FSE 2020 **MTFuzz: fuzzing with a multi-task neural network**

Dongdong She, Rahul Krishna, **Lu Yan**, Suman Jana, Baishakhi Ray

TL;DR: We propose MTFuzz, a fuzzing framework that improves code coverage using a multi-task neural network (MTNN). The main task focuses on edge coverage, while auxiliary tasks predict approach-sensitive and context-sensitive coverage. MTFuzz achieves up to 3× edge coverage and discovering 11 new bugs across 10 real-world programs.

▸IEEE CNS 2019 **Dynamic traffic feature camouflaging via generative adversarial networks**

Jie Li, Lu Zhou, Huaxin Li, **Lu Yan**, Haojin Zhu

TL;DR: We present FlowGAN, the first work to leverage GANs for dynamic traffic camouflaging to protect against traffic analysis attacks. FlowGAN morphs censored network traffic to resemble permitted flows, ensuring indistinguishability and preventing censorship. Evaluated on 10,000 real-world traffic flows, FlowGAN achieves 95%+ AUC within 0.5s.

▸J.SOFTW. 2023 **CrossFix: Resolution of GitHub issues via similar bugs recommendation**

Shin Hwei Tan, Ziqiang Li, **Lu Yan**

TL;DR: CrossFix is the first collaborative bug fixing tool by recommending similar GitHub issues, based on AST-based code comparison, Jaccard similarity on dependencies, and tree edit distance for Android UI components. CrossFix assists in bug fixes in 25% of the cases, and provides useful context in 55.56% of cases, with positive developer feedback.

Professional Experience

Tencent Keen Security Lab *March 2020 - May 2020*
Research Intern *Shanghai*

- Developed Seq2Seq and SeqGAN with real-time feedback to generate PNG inputs and improve code coverage on libpng.
- Evaluated the generated inputs by comparing fuzzing outcomes with random inputs and analyzing source code coverage via LLVM.
- Integrated Continuous Integration (CI) to automate fuzzing runs and result reporting.

Awards and Honors

Usenix Security Travel Grant	*2024*
NeurIPS Travel Grant	*2023*
Hongyi Scholarship (top 10 in overseas research, Shanghai Jiao Tong University)	*2019*
Zhiyuan Honor Scholarship (awarded for academic excellence in Shanghai Jiao Tong University)	*2016-2019*

Services

Reviewer: ICLR 2025, Journal of Computer Networks 2024.

External Reviewer: CCS 2025, Usenix Security 2024-2025, ICLR 2024, ICML 2023, NeurIPS 2023, theWebConf 2024.

Teaching Assistant: CS 580 Algorithm, Spring & Fall 2022, Purdue University.

Misc.: Assistant in hands-on science projects at summer camp at Blackfeet reservation, MT, June 2023 (media coverage)