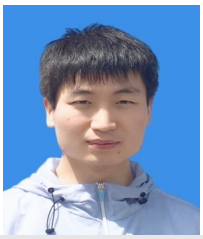
Yige Li

1995/03/26 (*BirthDay*)

Research Fellow, Singapore Management University xdliyige@gmail.com (E-Mail)

Google Scholar https://github.com/bboylyg



About Me

- I pursue research in **Trustworthy AI**, aiming to build secure, robust, and interpretable systems that align with human values and cognition. I' m especially interested in **Generative Models/Agent and AI safety**, and I seek simple yet insightful solutions grounded in theory. Guided by the philosophy **"Everything should be made as simple as possible, but not simpler,"** I approach research with both rigor and curiosity. Outside of work, I enjoy **rock climbing** and **swimming**.

Education Background

- **Xidian University** *Computer Science* 2019.09 – 2023.09 Doctor Degree
  - Trustworthy ML
- **Dalian Jiaotong University** *Control Engineering* 2017.09 – 2019.06 Master Degree
  - Pattern Recognition
- **Shijiazhuang Railway University** *Electrical Engineering* 2013.09 – 2017.06 Bachelor Degree
  - Data Structures, Automatic Control

Work Experience

- **Singapore Management University** *Computer Science* 2024.02 – up to now Postdoc Fellow
  - Trustworthy ML

Research Interest

- AI Safety on LLMs, VLMs and Agents
- Jailbreaking/backdoor attacks or defenses on agentic models
- Mitigating hallucinations and improving agentic model reliability

Award Honors

- **BackdoorLLM** won the **First Prize** in the SafetyBench competition organized by the Center for AI Safety.
- Outstanding doctoral candidate
- Tencent Academic second-class Scholarship

Internship Experience

- **Fudan University** *Research Intern* 2023.10 – 2024.01
  - Trustworthy ML
- **Shanghai AI Lab** *Research Intern* 2023.6 – 2023.08
  - Ethics in ML

- **Bosch R&D Center in Shanghai** *Computer Vision (Intern)* 2022.10 – 2023.02
  - Model adaptive training
  - Research on safety technology

Representative Work (* Corresponding Author)

- **Yige Li**, Xixiang Lyu, Xingjun Ma, Nodens Koren, Lingjuan Lyu, Bo Li, Yu-Gang Jiang, "Reconstructive Neuron Pruning for Backdoor Defense" , **ICML** 2023
- **Yige Li**, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, Xingjun Ma, "Anti-Backdoor Learning: Training Clean Models on Poisoned Data" , **NeurIPS** 2021.
- **Yige Li**, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, Xingjun Ma, "Neural Attention Distillation: Erasing Backdoor Triggers from Deep Neural Networks" , **ICLR** 2021.
- **Yige Li**, Hanxun Huang, Yunhan Zhao, Xingjun Ma, Jun Sun, "BackdoorLLM: A Comprehensive Benchmark for Backdoor Attacks and Defenses on Large Language Models", **First Prize** in the SafetyBench Competition 2025
- **Yige Li**, Jiabo He, Hanxun Huang, Jun Sun, Xingjun Ma, "Shortcuts Everywhere and Nowhere: Exploring Multi-Trigger Backdoor Attacks", Accepted at **IEEE TDSC**
- Hanxun Huang, SM Erfani, **Yige Li***, Xingjun Ma, James Bailey, "Detecting Poisoning Backdoor Attacks on CLIP" , **ICLR** 2025
- Wei Zhao, Ze Li, **Yige Li**, Jun Sun, "Q-MLLM: Vector Quantization for Robust Multimodal Large Language Model Security" , **NDSS** 2025.
- Nay Myat Min, Long H Pham, **Yige Li***, Jun Sun, "CROW: Eliminating Backdoors from Large Language Models via Internal Consistency Regularization" , **ICML** 2025
- Hanxun Huang, SM Erfani, **Yige Li***, Xingjun Ma, James Bailey, "X-Transfer Attacks: Towards Super Transferable Adversarial Attacks on CLIP" , **ICML** 2025
- Jiaming Zhang, Junhong Ye, Xingjun Ma, **Yige Li***, Yunfan Yang, Jitao Sang, Dit-Yan Yeung, "Anyattack: Towards large-scale self-supervised generation of targeted adversarial examples for vision-language models" , **CVPR** 2025
- Peihai Jiang, Xixiang Lyu, **Yige Li***, Jing Ma, "Backdoor Token Unlearning: Exposing and Defending Backdoors in Pretrained Language Models" , **AAAI** 2025.
- **Yige Li**, Peihai Jiang, Jun Sun, Peng Shu, Tianming Liu, Zhen Xiang, "Adaptive Content Restriction for Large Language Models via Suffix Optimization", Under review at **NeurIPS** 2025
- **Yige Li**, Hanxun Huang, Jiaming Zhang, Xingjun Ma, Yu-Gang Jiang, "Expose Before You Defend: Better Backdoor Defense With Exposed Models", Under Review at **AAAI** 2026
- Shen Dong, Shaochen Xu, Pengfei He, **Yige Li**, Jiliang Tang, Tianming Liu, Hui Liu, Zhen Xiang, "A Practical Memory Injection Attack against LLM Agents", Under Review at **NeurIPS** 2025
- Yunhan Zhao, Xiang Zheng, Lin Luo, **Yige Li**, Xingjun Ma, Yu-Gang Jiang, BlueSuffix: Reinforced Blue Teaming for Vision-Language Models Against Jailbreak Attacks" , **ICLR** 2025.
- Wei Zhao, Ze Li, **Yige Li**, Ye Zhang, Jun Sun, "Defending Large Language Models Against Jailbreak Attacks via Layer-specific Editing" , **EMNLP** 2024.
- More publications are available on my Google Scholar