Hou Guanyu (侯冠宇)

Undergraduate student majoring in Software engineering

📍 Chengdu, Sichuan, China | ☎ +86 19934322578 | @ edgarhou03@gmail.com

Education

Chengdu University of Technology **September 2021 - Present**

Software Engineering Sino-British Collaborative Education (with Oxford Brookes University)

Main modules (based on the UK undergraduate grading system)

- Object Orientated Programming 90% (using Java)
- Problem Solving and Programming 89% (using Python)
- Foundation of Security 83%

Experience

Data Stealing Attacks against Large Language Models via Backdooring **May 2024 - July 2024**

Corresponding author

🔗 https://www.mdpi.com/2079-9292/13/14/2858

- Produce fine-tuned datasets and write prompts to train and test the attacked model including GPT 3.5 and Mistral 7B and validate the attack performance. This paper has already accepted for ***Electronic***.
- Designing and running experiments, collecting experimental data, and using this to plot figures and create tables to show how our methods perform under different conditions.
- Wrote the Experimental Results section, which describes the experimental data in detail and explains its significance.

Deep Learning, Privacy, LLM

Embedding based Sensitive element injection against Text-to-Image Generative Models **Marth 2024 - April 2024**

Co-first author

- Innovated an attack on txt2img models, misleading them to generate images with sensitive elements. Pioneered a new vector for assessing AI model security. This paper has already accepted for **ICSP 2024**.
- Authored methods and conducted experiments on model susceptibility to varying degrees of attack. Advanced the understanding of model robustness and vulnerabilities.
- Adapted a VGG16 model for binary classification of sensitive content in generated images. Contributed a practical tool for evaluating AI-generated content safety.

text-to-image, NLP

Talk Too Much: Poisoning Large Language Models under Token Limit **January 2024 - Marth 2024**

Co-auther

🔗 https://arxiv.org/abs/2404.14795

- Fine-tuned advanced language models, **GPT-3.5** and **Mistral**, validating the proposed methods.
- Produce fine-tuned datasets and write prompts to train and test the attacked model and validate the attack.
- Devised comprehensive experimental frameworks, encompassing both ablation and comparative studies, to substantiate the paper’s hypotheses and validate the effectiveness of the techniques introduced.

Large Language Model, Fine-tune, Deep Learning

Skills

Programming Languages	**Deep Learning Frameworks**	**Data Visualization Tool**
Python, Java, C/C++	Pytorch	Matlab