Guanyu HOU

D.O.B: 24/06/2003 **Tel.:** 86-19934322578

Location: Sichuan, China **Email:** edgarhou03@gmail.com

ACADEMIC BACKGROUND

Chinese-Foreign Cooperation in Running Schools (4+0) *09/2021-06/2025*

Chengdu University of Technology, China

Oxford Brookes College

✧ **Major:** Software Engineering; **Degree:** Bachelor's Degree in Engineering(Pending)

✧ **GPA:** 3.12/4.00

Oxford Brookes University, UK

✧ **Major:** Software Engineering; **Degree:** Bachelor of Science (Expected to be conferred with a First Class degree)

✧ **AVG:** 74.994/100

✧ **Medium of Instruction:** Chinese & English

PAPER & PUBLICATION

Data Stealing Attacks against Large Language Models via Backdooring

Jiaming He, **Guanyu Hou***, Xinyue Jia, Yangyang Chen, Wenqi Liao, Yinhang Zhou, Rang Zhou (* Corresponding Author)

✧ Published in the ***Electronics*** (2024, 13(14), 2858) on 19 July 2024

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

Responsibilities:

✧ Developed experimental codes, created backdoor-poisoned datasets, simulating a knowledge base for cheat sheet in the production environment using Pinecone

✧ Implemented fine-tuning strategies via OpenAI and LoRA to conduct attack experiments and ablation experiments under varied parameter conditions

✧ Evaluated the FastKASSIM, cosine similarity and ASR between the data from the poisoned model and the original data through three different types of benchmarking

✧ Collected experimental data, created all necessary figures using MATLAB, provided clear descriptions to demonstrate attack effects and examples under different parameter settings, and compared results with that of PLeak

✧ Also acted as the corresponding author and supervised and reported progress to the first author and the advisor

Embedding Based Sensitive Element Injection against Text-to-Image Generative Models

Benrui Jiang*, Kan Chen*, **Guanyu Hou***, Xiying Chen*, Jiaming He (* Equally Contribution)

✧ Accepted by the ***2024 9th International Conference on Intelligent Computing and Signal Processing (ICSP***) on 17 April 2024, and will be published in IEEE (ISBN: 979-8-3503-7654-8) and submitted for index in IEEE Xplore, EI Compendex, and Scopus afterward

✧ Link: https://orangestella.github.io/res/ICSP.pdf

Responsibilities:

✧ Embedded the prepared poisoned word embeddings into text embeddings generated by the Prompt Encoder

✧ Used MATLAB to create all figures and their detailed descriptions for the paper so as to show attack effects and examples clearly

✧ Authored the methodology section, and provided detailed explanations and equations to clarify all employed research methods

Talk Too Much: Poisoning Large Language Models under Token Limit

Jiaming He, Wenbo Jiang, **Guanyu Hou**, Wenshu Fan, Rui Zhang, Hongwei Li

✧ Submitted to the arXiv on 23 April 2024, and will be submitted to AAAI Conference on Artificial Intelligence in August 2024 for further publication

✧ Link: https://arxiv.org/abs/2404.14795

Responsibilities:

✧ Wrote the experimental codes and prepared the poisoned datasets, conducted the attack experiments and ablation experiments under various parameter conditions using OpenAI and SFT (Supervised Fine-Tuning), and utilized FastEval for diverse evaluations of poisoned models, including Chain of Thought

- Gathered the experimental data and created figures for this paper using MATLAB, including graphs and tables, to clearly show the effects and examples of attacks
- Brought up the idea of Poison Agent, a tool for generating poisoned data, and successfully implemented it with GPT-4-turbo in the OpenAI API

INTERNSHIP

Software Testing Intern | **Chengdu Allition Technology Co., Ltd** *15/06/2024-25/07/2024*

- Developed a great deal of test cases and wrote robust unit tests with JUnit 4 to conduct functional testing on the target software
- Performed black-box, white-box, and regression testing using tools like Postman
- Consistently identified an average of 5 bugs or failures daily, with approximately 15% of test cases successfully detecting issues

COMPUTER SKILLS

- **Languages:** Python (3 years), Java (2 years), SQL(2 years), and C/C++ (1 year)
- **AI/ML:** NumPy (1 year) and PyTorch (1 year)
- **Miscellaneous:** Pinecone(1 year) and MATLAB (1 year)