

Guanyu Hou
Trustworthy AI
edgarhou03@gmail.com | Website

RESEARCH INTEREST

My research interest lies in large language models and multimodal systems. I focus on building capable and reliable deep learning models across language and vision. I have worked extensively on model robustness, red-teaming, and data privacy in large language models and multimodal systems.

EDUCATION

Master of Science in Artificial Intelligence

Year of completion: 2026 Expected
The University of Manchester

Bachelor of Science with First Class Honors in Software Engineering

Year of completion: 2025
Oxford Brookes University

INTERNSHIP EXPERIENCE

Shanghai Artificial Intelligence Laboratory (PJLab)

Research Intern (Advisor: Dr. Lijun Li)

Shanghai, China
Sept 2026 – Present

- Focus on AI4Science safety, exploring vulnerability detection, safety boundaries, and alignment mechanisms for scientific domains.

ByteDance

Algorithm Intern (Intelligent Security)

Beijing, China
Jul 2026 – Sept 2026

- **Multilingual Safety Classifier:** Deployed a LLM firewall classifier via SFT and data augmentation, achieving **>97% Precision/Recall/F1** globally.
- **Long-Reasoning Optimization:** Devised preprocessing and augmentation pipelines for complex reasoning inputs, driving a **+1% P/R improvement** in internal training infra.
- **Jailbreak Defense:** Synthesized targeted hard cases from red-team logs using agent-based mining, significantly enhancing robustness against prompt injections.

PUBLICATIONS

[EMNLP 2025] (CCF-B) Evaluating Robustness of Large Audio Language Models to Audio Injection: An Empirical Study. (*Accepted 2025.7*)

Guanyu Hou*, Jiaming He* (* *Equal contribution*), Yinhang Zhou, Ji Guo, Yitong Qiao, Rui Zhang, Wenbo Jiang

[CVPR 2026] (CCF-A) TEAR: Temporal-aware Automated Red-teaming for Text-to-Video Models. (*Accepted 2026.2*) (**Oral, 0.88%**)

Jiaming He, **Guanyu Hou**, Hongwei Li, Zhicong Huang, Kangjie Chen, Yi Yu, Wenbo Jiang, Guowen Xu, Tianwei Zhang

[ICASSP 2025] (CCF-B) PRESS: Defending Privacy in Retrieval-Augmented Generation via Embedding Space Shifting. (*Accepted 2024.12*)

Jiaming He*, Cheng Liu*, **Guanyu Hou*** (* *Equal contribution*), Wenbo Jiang, Jiachen Li

[MM Asia 2025] (CCF-C) When Hallucinated Concepts Cross Modals: Unveiling Backdoor Vulnerability in Multi-modal In-context Learning. (*Accepted 2025.12*)

Guanyu Hou, Jiaming He, Yitong Qiao, Jiachen Li, Zihan Wang, Qiyang Song, Hongwei Li, Wenbo Jiang

[AAAI 2025] (CCF-A) Watch Out for Your Guidance on Generation! Exploring Conditional Backdoor Attacks against Large Language Models. (*Accepted 2024.11*) (**Oral, 4.6%**)

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

[ICME 2025] (CCF-B) Weaponizing Tokens: Backdooring Text-to-Image Generation via Token Remapping. (*Accepted 2025.3*)

Jiaming He, Wenbo Jiang, **Guanyu Hou**, Qiyang Song, Guo Ji and Hongwei Li

[**Electronics**] (**JCR Q2**) Data Stealing Attacks against Large Language Models via Backdooring. (*Accepted 2024.7*)

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

[**AAAI 2027**] (**CCF-A**) ChronoLock: Protecting Videos from Unauthorized Text-to-Video Personalization. (*Under review*)

Jiaming He, Jiashu Zhang, **Guanyu Hou**, Shuhan Ye, Hanwei Zhu, Yi Yu, Xudong Jiang

ACADEMIC SERVICE

- AAAI 2026 Reviewer
- NeurIPS 2026 Reviewer
- AAAI 2027 Program Committee Member

LANGUAGES

English

IELTS 7.0

SKILLS & TOOLS

ML Frameworks and Tools

ML frameworks: PyTorch, Scikit-learn

Tools: Hugging Face (Transformers, PEFT, TRL), Llama-Factory

Software Engineering

Programming Languages: Python, Java, SQL

DevOps: Git, Linux/Shell