WENJIE FU

+86-188-1110-6219 | wjfu99@outlook.com | https://wjfu99.github.io/
wjfu99 | Wenjie Fu | Wenjie Fu[0005]
Wuhan, Hubei - 430074, China

BIO

I am a fourth-year Ph.D. student at Huazhong University of Science and Technology, under the supervision of Prof. Tao Jiang. I am also a long-term research intern at the FIB Lab, Tsinghua University, co-advised by Prof. Yong Li. My current research interest lies in the area of AI Privacy & Security. Specifically, I am interested in the study of membership inference attacks and extraction attacks, with a recent focus on the vulnerability of large language models (LLMs) and generative models. I am also committed to exploring potential remedies against these threats and jaillbreak attaks. Before that, I had invesitigated the privacy-perserving algorithms in the field of mobile big data mining for a while.

EDUCATION

- **B.E., Beijing Jiaotong University (BJTU)** *Sep, 2017 - Jun, 2021*
  School of Electronics and Information Engineering Beijing, China
  - **Major**: Telecommunication Engineering GPA: 3.96/4.00
  - **Bachelor Thesis**: "Research on Offloading Algorithm of Mobile Edge Computing Based on Machine Learning"
- **Ph.D Student, Huahzong University of Science and Technology (HUST)** *Sep, 2021 - Present*
  Wuhan National Laboratory for Optoelectronics (Sep, 2021 - Feb, 2023)
  School of Cyper Science and Engineering (March, 2023 - Present) Wuhan, China
  - **Major**: Information and Communication Engineering GPA: 3.58/4.00
  - **Research Interests**: Trustworth AI, Privacy & Security, Large Language Model and Data Mining
  - **Supervisor**: Prof. Tao Jiang

EXPERIENCE

- **Exchange Student, National Chiao Tung University (NCTU)** *Sep, 2019 - Jan, 2020*
  Applied Computing and Multimedia Lab, Dept. of CS Hsinchu, Taiwan
  - **Major**: Computer Science GPA: 4.00/4.00
  - **Project**: Continual-level Image Restroation
  - **Supervisor**: Prof. Ching-Chun Huang
- **Researcher&Intern, Tsinghua University (THU)** *Dec, 2021 - Present*
  Visiting Researcher in Future Intelligence Lab, Dept. of EE (Dec, 2021 - March, 2022) Beijing, China
  Remote Intern in Future Intelligence Lab, Dept. of EE (Mar, 2021 - Present)
  - **Major**: Information and Communication Engineering
  - **Project**: Urban Epidemic Simulator (Based on UE4 engine)
  - **Supervisor**: Prof. Yong Li

PATENTS AND PUBLICATIONS

*=EQUAL CONTRIBUTION, C=CONFERENCE, J=JOURNAL, S=IN SUBMISSION, P=PATENT

[C.1] **Fu, W.**, Wang, H., Gao, C., Liu, G., Li, Y., & Jiang, T. (2024). **Membership Inference Attacks against Fine-tuned Large Language Models via Self-prompt Calibration.** *In The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*. [Paper] [Code] [Slides]

[C.2] **Fu, W.**, Wang, H., Gao, C., Liu, G., Li, Y., & Jiang, T. (2025). **MIA-Tuner: Adapting Large Language Models as Pre-training Text Detector.** *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI, Oral)*. [Paper] [Code] [Slides]

[S.1] **Fu, W.**, Wang, H., Gao, C., Liu, G., Li, Y., & Jiang, T. (2023). **A Probabilistic Fluctuation based Membership Inference Attack for Diffusion Models.** *arXiv preprint arXiv:2308.12143*. (Submitted to IEEE TIFS) [Paper] [Code]

[S.2] Wang, H., **Fu, W.***, Tang, Y., Chen, Z., Huang, Y., Piao, J., Gao, C., Xu, F., Jiang, T., & Li, Y. (2025). **A Survey on Responsible LLMs: Inherent Risk, Malicious Use, and Mitigation Strategy.** *arXiv preprint* arXiv:2501.09431. (Submitted to ACM CSUR) [Paper]

[S.3] Deng, H., Tang, Y., **Fu, W.**, Wang, H., Chen, K., & Jiang, T. (2025). **FedSkeleton: Secure Multi-Party Graph Skeleton Construction for Privacy-Preserving Federated Time-Series Forecasting.** (Submitted to KDD'25)

[J.1] **Fu, W.**, Wang, H., Gao, C., Liu, G., Li, Y., & Jiang, T. (2024). **Privacy-Preserving Individual-Level COVID-19 Infection Prediction via Federated Graph Learning.** *ACM Transactions on Information Systems*, 42(3), 1-29. [Paper] [Code]

PROJECTS

1. **Memorization-based Attacks and Defense in Large Language Models (LLMs)** *Jul, 2023 - Present*
National Key Research and Development Project THU

➢ **Project Objective**: Develop memorization-based attacks that achieve a success rate of over 90% against LLMs and design robust safeguards to counter them.

- Propose a MIA method based on Self-calibrated Probabilistic Variation for fine-tuned LLMs, where I propose a self-prompt approach to extract reference dataset from LLM itself in a practical manner, then introduce a more reliable membership signal based on memorization rather than overfitting.. (*NeurIPS Paper*)
- Collecte and release a more up-to-date dataset, WIKIMIA-24, for evaluating MIAs against pre-trained LLMs. (*AAAI Oral Paper*)
- Design a novel MIA method that can persuade pre-trained LLMs themselves to serve as effective and efficient attackers. Two instances of MIA-Tuner can be applied to both aligned and unaligned LLMs. (*AAAI Oral*)
- Develop two safeguards that can reduce the accuracy of MIA to that of a random guesser, without compromising the linguistic quality of the LLMs. (*AAAI Oral Paper*)

2. **Multi-Modal Deepfake Detection and Trace Removal** *Mar, 2022 - Present*
Frontier Technology Innovation Program THU

➢ **Project Objective**: Develop a deepfake detection system based for AIGC models across multi-modal, including text, image, audio, and video. Then propose corresponding trace removal mechanisms to decrease the detection accuracy.

- Independently design a efficient and lightweight detection algorithm based on the perturbation mechanism for LLM-generated texts. Undertook the deployment and the self-evaluation of the LLM-generated text detection system. Participated the design, deployment and self-evaluation of the model-generated image detection system.
- Develop a post-processing and resampling pipeline for removing the generative trace in LLM-generated texts.

3. **Detecting Training Data of Generative Models through the Lens of Memorization** *Nov, 2022 - Jul, 2023*
National Key Research and Development Project Subject THU

➢ **Project Objective**: Detect the training data of generative models through a black-box API access.

- Reveal and verify the phenomenon that existing MIA algorithms largely rely on overfitting in generative models, which can be avoided by several regularization methods.
- Present a Probabilistic Fluctuation Assessing Membership Inference Attack (PFAMI) based on the distinct probabilistic fluctuation characteristics of members and non-members.

4. **Infectious Disease Forecasting Based on Urban Mobility Network** *Jan, 2022 - Dec, 2025*
Joint Fund Project HUST & THU

➢ **Project Objective**: Establish algorithmic models for epidemic forecasting, policy formulation, and intelligent decision support.

- Primary concentrate on leveraging mobility data for infection case detection while ensuring user privacy. Design a novel spatio-temporal hypergraph construction method for detection and incorporate a obfuscation mechanism to protect user privacy.
- Investigate the individual-level infection prediction for more precise individual-level intervention strategies (e.g., early warning and mobility control) and propose Falcon, a privacy-preserving federated graph learning framework. (*TOIS Paper*)

5. **Urban Epidemic Simulator** *Dec, 2021 - Feb, 2023*
National 5G+ Medical and Health Application Pilot Project HUST & THU

➢ **Project Objective**: Achieve individual-level infectious disease transmission 3D spatiotemporal visualization, alerting on infection pathways and weak points in prevention and control through visualization, providing references for prevention and control decision-making.

- Independently undertook the development of the human mobility simulation and the epidemic spread simulation modules; participated in the construction of the database and the deployment of the data query API.
- Accomplish the construction of 3D building models and the extraction of road networks in the Wuhan city. Generate second-level fine-grained trajectories based on Original-Destination mobility data.

SKILLS

- **Programming Languages:** Python, C, C++, MATLAB, Javascript, Fortran, Assembly
- **LLM Related Packages:** Pytorch, TensorFlow, Accelerate, DeepSpeed, Transformers, PEFT, TRL
- **Database Systems:** MySQL, Django, MongoDB

SERVICES

- **Conference Reviewer**
  - NeurIPS (Conference on Neural Information Processing Systems)
  - AAAI (AAAI Conference on Artificial Intelligence)
  - KDD (SIGKDD Conference on Knowledge Discovery and Data Mining)
  - TheWebConf/WWW (The Web Conference)
- **Journal Reviewer**
  - IoTJ (IEEE Internet of Things Journal)
  - SCIS (Science China-Information Sciences)

Honors and Awards

1. **Third-class Scholarship for Doctoral Students** *2024 - 2025*
   School of Cyper Science and Engineering, HUST Wuhan, China
2. **Fisrt-class Scholarship for Doctoral Students** *2023 - 2024*
   School of Cyper Science and Engineering, HUST Wuhan, China
3. **Fisrt-class Scholarship for Doctoral Students** *2022 - 2023*
   School of Cyper Science and Engineering, HUST Wuhan, China
4. **Fisrt-class Scholarship for Doctoral Students** *2021 - 2022*
   School of Cyper Science and Engineering, HUST Wuhan, China
5. **First-class Academic Excellence Scholarship** *2020 - 2021*
   School of Electronics and Information Engineering, BJTU Beijing, China
6. **National Encouragement Scholarship** *2018 - 2019*
   School of Electronics and Information Engineering, BJTU Beijing, China
7. **Second-class Academic Excellence Scholarship** *2018 - 2019*
   School of Electronics and Information Engineering, BJTU Beijing, China
8. **Second Prize, 29th Beijing College Student Mathematics Competition** *Nov, 2018*
   Beijing Mathematical Society (BMS) Beijing, China
9. **Second Prize, 10th National College Mathematics Competition** *Nov, 2018*
   Popularization Committee of Chinese Mathematical Society Beijing, China
10. **Outstanding Social Work Scholarship** *2018 - 2019*
    School of Electronics and Information Engineering, BJTU Beijing, China
11. **Second Class Academic Excellence Scholarship** *2017 - 2018*
    School of Electronics and Information Engineering, BJTU Beijing, China