Yisheng Zhong

yzhong7@gmu.edu | easonzhong99.github.io | in/yisheng-zhong

Summary

PhD student in Cybersecurity at **George Mason University** (advised by Dr. Zhuangdi Zhu). Research focuses on the security and privacy of large language models (LLMs), including **unlearning**, **alignment**, and **defenses against LLM-driven content extraction and misuse**. Master's at the University of Chinese Academy of Sciences with work on privacy-preserving federated learning at the State Key Laboratory of Information Security.

Research Interests

- Security & privacy of LLMs: LLM Unlearning, Safety Alignment, and Defense Against Unauthorized Retrieval,
- Privacy-preserving & Byzantine-robust Federated Learning

Education

George Mason University Fall 2024 – Present
Ph.D. of Information Technology (Cybersecurity) *GPA: 4.00/4.00*

- Research on security and privacy of LLMs, with a particular interest in **LLM unlearning**.
- Graduate Research Assistant (2024). Graduate Teaching Assistant (2025)

University of Chinese Academy of Sciences Fall 2021 – Spring 2024
Master of Cyber Security *GPA: 3.56/4.00*

- Relevant coursework: Machine Learning, Deep Learning, Security Protocols, Applied Cryptography.
- Focus on privacy-preserving computing and defenses against adversarial attacks in machine learning.

Harbin University of Science and Technology Fall 2017 – Spring 2021
Bachelor of Computer Science *GPA: 3.84/4.00 (Top 1%)*

- Relevant coursework: Advanced Mathematics, Data Structure, Discrete Mathematics, Probability & Statistics, Linear Algebra, Pattern Recognition
- Received **direct admission offer** to pursue a master's degree at the Chinese Academy of Sciences.

Research Experience

1. **DUET: Distilled LLM Unlearning from an Efficiently Contextualized Teacher** *Under review, ICLR 2026*
   Yisheng Zhong, Zhengbang Yang, Zhuangdi Zhu
   - Developed DUET, an LLM unlearning technique that (i) uses an efficiently contextualized teacher (prompt-conditioned) to **demonstrate refusals on undesirable knowledge** and (ii) **distills this behavior into a student model**, achieving targeted forgetting with minimal utility loss.
   - Introduced **Top-K logit alignment** in place of full-vocabulary KL, enabling more precise forgetting with better efficiency; on MUSE, average leakage decreased by 4% (ROUGE-Forget) and utility increased by 10% (ROUGE-Retain/MMLU), while training used about 1/645 of corpus tokens (2,233 vs. 1.44M).
   - Demonstrated robustness to reverse prompts and task-format shift (QA $\rightarrow$ continuation) on WMDP and MUSE; under reverse prompts, the in-context teacher's leakage rises from 4.52% to 37.62%, whereas DUET remains around 5.98%$\rightarrow$7.27%.
2. **Web Intellectual Property at Risk: Preventing Unauthorized Real-Time Retrieval by Large Language Models** *EMNLP 2025, Main Conference*
   Yisheng Zhong, Yizhu Wen, Junfeng Guo, Mehran Kafai, Heng Huang, Hanqing Guo, Zhuangdi Zhu
   - Designed a **semantic defense framework** that embeds optimized HTML policy cues to block LLMs from unauthorized real-time extraction and redistribution of web content.

- Improved defense success rates from 2.5% to 88.6% across multiple proprietary LLMs, outperforming configuration-based defenses such as robots.txt.
- Supported three granular protection goals: refusal to answer, partial masking, and redirection to the source, ensuring both autonomy and content discoverability.

3. **Hierarchical Federated Unlearning for Large Language Models** *FedKDD 2025*
   Yisheng Zhong, Zhengbang Yang, Zhuangdi Zhu
   - Proposed **Federated UnLearning Merge (FULM)**, a scalable and privacy-preserving framework for decentralized LLM unlearning requests.
   - Decoupled forgetting and retention into dual task adapters and introduced a **hierarchical merging strategy** to mitigate inter- and intra-domain interference.
   - Demonstrated effectiveness on WMDP, TOFU, and MUSE benchmarks, showing improved trade-offs between forgetting performance and retention utility; Overall improved by 8.9% on TOFU and 7.8% on the heterogeneous WMDP+MUSE setting.
4. **PROFL: A Privacy-Preserving Federated Learning Method with Stringent Defense Against Poisoning Attacks** *CSCWD 2023*
   Yisheng Zhong, Li-Ping Wang
   - Developed a **Byzantine-robust federated learning framework** combining similarity-based and statistical defenses against hidden data poisoning.
   - Integrated two-trapdoor Homomorphic Encryption for secure computation; outperformed comparable schemes in extreme cases by 13%–56%.
5. **CATNIP: LLM Unlearning via Calibrated and Tokenized Negative Preference Alignment** *Under review, ICLR 2026*
   Zhengbang Yang, Yisheng Zhong, Junyuan Hong, Zhuangdi Zhu
   - Propose CATNIP, a **retention-data-free unlearning objective** that calibrates gradient updates with an adaptive reverse-policy reference and applies token-level weighting to target high-confidence tokens while limiting collateral damage.
   - Demonstrate stronger forget–retain trade-offs than methods GA/NPO/SimNPO/FLAT on benchmark WMDP (Bio/Cyber) and MUSE-Books; remains robust with scarce, short QA-format unlearning data.
   - Ablations isolate gains from calibration and tokenization, showing both components are necessary to achieve effective forgetting without retention or contrastive pairs.

Competition Experience

Mathematical Contest in Modeling (MCM) — *Meritorious Winner* (International First Prize) Apr 2020
Team leader
- Modeled a 3D cellular automata system governed by differential equations; solved via a genetic algorithm.
- Led pre-competition preparation, coordinated team roles, and oversaw algorithm design, optimization, and implementation.

ACADEMIC SERVICES

ICLR 2024 and 2025 Reviewer

Honors & Awards

Meritorious Winner at the Mathematical Contest in Modeling (**First Prize**), 2020

Annual Scholarships of the Chinese Academy of Sciences, 2021–2023

First-Class Scholarships of Harbin University of Science and Technology, 2017–2021