Zhihao Zhu

Postdoctoral Fellow at HKUST

zhihaozhu@ust.hk GitHub Google Scholar ORCID

Experience

- **Postdoctoral Fellow**, Hong Kong University of Science and Technology, advised by Prof. Yi Yang.
- **Ph.D. in Computer Science**, University of Science and Technology of China (2025), advised by Prof. Defu Lian.
- **Bachelor's degree in Computer Science**, Fuzhou University (2020).

Research Focus

I study trustworthy AI and large language models, focusing on auditing, privacy, and training data exposure.

Trustworthy AI

Auditing model behavior, data revocation, membership inference, and model stealing risks in recommender and graph learning systems.

Large Language Models

Detecting training data exposure and understanding privacy risks in language and vision-language models, emphasizing pre-training data identification and leakage measurement.

Published Journal Article

Forget Me If You Can: Auditing User Data Revocation in Recommendation Systems
Zhihao Zhu, Yi Yang, Yangyang Fan, Defu Lian
Information Systems Research 2026

Working Papers

HoneyImage: Verifiable, Harmless, and Stealthy Dataset Ownership Verification for Image Models
Zhihao Zhu, Jiale Han, Yi Yang
Management Information Systems Quarterly Major Revision

Revealing Training Data Exposure in Vision-Language Large Models via Parameter Gradients
Zhihao Zhu, Hongyi Tang, Yi Yang, Ahmed Abbasi
Nature Communications Major Revision

RecShield: Output-Level Attribute Unlearning in Recommender Systems
Zhihao Zhu, Yi Yang
Information Systems Research Under Review

GraphMSA: Stress Testing Graph Classification Services Against Model Stealing Attacks
Zhihao Zhu, Yi Yang, Chenwang Wu, Defu Lian
INFORMS Journal on Computing Under Review

AI Conferences

TDDBench: A Benchmark for Training Data Detection
Zhihao Zhu, Yi Yang, Defu Lian
International Conference on Learning Representations 2025

Identifying Pre-training Data in LLMs: A Neuron Activation-Based Detection Framework
Hongyi Tang*, **Zhihao Zhu***, Yi Yang. Equal contribution.
Conference on Empirical Methods in Natural Language Processing 2025

Membership Inference Attacks against Sequential Recommender Systems
Zhihao Zhu, Chenwang Wu, Rui Fan, Defu Lian, Enhong Chen
The Web Conference 2023

Service

Reviewer for *Information Systems Research* (ISR), *INFORMS Journal on Computing* (IJOC), NeurIPS, ICLR, and ACL Rolling Review.