

Yao Tong

tongyao@u.nus.edu • [Homepage](#) • [Google Scholar](#)

Research Interests

My research interests focus on understanding the capabilities and behaviors of models and mitigating catastrophic risks in AI. Recently, my work has centered on:

1. **Evaluating and understanding LLM behaviors:** hallucination, memorization, and extrapolative generalization
2. **Copyright protection:** developing verification methods for private data, model-generated works, and model architectures

Key words: Trustworthy machine learning; copyright for data and models; privacy auditing; memorization and generalization; ai security.

Education

Ph.D. in Computer Science

2022 – Present

National University of Singapore • Singapore

- Advisor: Prof. Reza Shokri.

B.S. in Computer Science

2018 – 2022

The Chinese University of Hong Kong • China

- Graduated with First Class Honours.

Preprints

2. **Decomposing Extrapolative Problem Solving: Spatial Transfer and Length Scaling with Map Worlds**

Yao Tong, Jiayuan Ye, Anastasia Borovykh, Reza Shokri

Under review (public on OpenReview), 2025

Extended version of our NeurIPS 2025 workshop paper below.

1. **Identifying Optimal Output Sets for Differential Privacy Auditing**

Yao Tong, Jiayuan Ye, Reza Shokri

Publications & Workshops

6. **SeedPrints: Fingerprints Can Even Tell Which Seed Your Large Language Model Was Trained From**

Yao Tong*, Haonan Wang*, Siquan Li, Kenji Kawaguchi, Tianyang Hu

In NeurIPS Workshop on Prevent Unauthorized Knowledge Use from Large Language Models, 2025

Full version on arXiv, under review

5. **When Transformers Can (or Can't) Generalize Compositionally? A Data-Distribution Perspective**

Yao Tong, Jiayuan Ye, Anastasia Borovykh, Reza Shokri

4. Cut the Deadwood Out: Training-Free Backdoor Purification via Guided Module Substitution

Yao Tong*, Weijun Li*, Xuanli He, Haolan Zhan, Qionghai Xu

In Findings of Association for Computational Linguistics EMNLP, 2025

3. How much of my dataset did you use? Quantitative Data Usage Inference in Machine Learning

Yao Tong*, Jiayuan Ye*, Sajjad Zarifzadeh, Reza Shokri

In International Conference of Learning Representations (ICLR), 2025

Oral Presentation (Top ~1.5% among submissions)

2. The Stronger the Diffusion Model, the Easier the Backdoor: Data Poisoning to Induce Copyright Breaches Without Adjusting Finetuning Pipeline

Haonan Wang, Qianli Shen, **Yao Tong**, Yang Zhang, Kenji Kawaguchi

In NeurIPS Workshop on Backdoors in Deep Learning, 2023

Oral Presentation

In International Conference on Machine Learning (ICML), 2024

Oral Presentation (Top ~2% among submissions)

1. Towards Regulatable AI Systems: Technical Gaps and Policy Opportunities

Xudong Shen, Hannah Brown, Jiashu Tao, Martin Strobel, **Yao Tong**, Akshay Narayan, Harold Soh, Finale Doshi-Velez

In Communications of the ACM (CACM), 2024

Work conducted during Finale's visit to NUS as an outcome of the RRAI Workshop; middle authors order is alphabetical.

* denotes equal contribution.

Selected Projects

Privacy Meter: An open-source library to audit data privacy in statistical and machine learning algorithm via membership inference. 2025

Open-source (**500+ stars**) • [GitHub](#)

- Implemented privacy auditing tools such as DUCI and RMIA.
- Contributed to the development and long-term maintenance of the library as one of the organizers.

Teaching

Teaching Assistant, CS5562 Trustworthy Machine Learning

National University of Singapore

2023 Fall

Teaching Assistant, CS3244 Machine Learning

National University of Singapore

2024 Spring

Teaching Assistant, CS6208 Advanced Topics in Artificial Intelligence

National University of Singapore

2024 Fall

Teaching Assistant, Data and Knowledge Management, Software Engineering <i>The Chinese University of Hong Kong</i>	2021 – 2022
--	-------------

Honors & Awards

Oral Paper Award - ICLR	2025
--------------------------------	------

Top Reviewer Award - NeurIPS	2025
-------------------------------------	------

Oral Paper Award - ICML	2024
--------------------------------	------

President Graduate Fellowship - NUS	2022 – Present
--	----------------

Dean's List - CUHK	2019 – 2022
---------------------------	-------------

University Research Award - CUHK	2021, 2022
---	------------

School Academic Scholarship (for Top 2% students) - CUHK	2020 – 2022
---	-------------

Bowen Scholarship - CUHK	2018 – 2022
---------------------------------	-------------

Service

Reviewer: NeurIPS 2025 (**Top Reviewer**), ICLR 2025, ICML Workshop 2025, NeurIPS Workshop 2025

Sub-reviewer: CCS 2024, USENIX Security 2024