Yao Tong

tongyao@u.nus.edu • Homepage • Google Scholar

Research Interests

My research focuses on understanding the behaviors and capabilities of large language models, the underlying mechanisms that give rise to them, and how these properties relate to and shape model trustworthiness. In particular, I am interested in:

1. **Model behaviors and capabilities.** Studying phenomena such as intrinsic biases and generalization, and understanding how they are shaped by different stages of the learning pipeline, including model initialization, data distributions, training paradigms (e.g., SFT and RL), and inference-time methods.
2. **Trustworthiness.** Studying how these properties relate to model trustworthiness: (i) as sources of security and privacy risks, and (ii) as signals that can be leveraged for detection and protection, such as model fingerprinting and data usage auditing.

More recently, I have also become interested in agentic memory and personalization systems, especially in how long-term memory and adaptive interaction introduce new failure modes, privacy risks, and auditing challenges.

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.

Experience

Visiting Researcher 2026 – Present

***ETH Zurich**, SPY Lab* • Zurich, Switzerland

- Host: Prof. **Florian Tramer**.
- Research on security and privacy in agentic memory systems.

Publications

6. **Generalization in LLM Problem Solving: The Case of the Shortest Path**
Yao Tong, Jiayuan Ye, Anastasia Borovykh, Reza Shokri
In International Conference of Learning Representations (ICLR), 2026
(Previously titled: Decomposing Extrapolative Problem Solving: Spatial Transfer and Length Scaling with Map Worlds)

5. **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 International Conference of Learning Representations (ICLR), 2026

4. **Cut the Deadwood Out: Training-Free Backdoor Purification via Guided Module Substitution**
   Yao Tong*, Weijun Li*, Xuanli He, Haolan Zhan, Qiongkai 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.*

Preprints & Workshops

2. **Transformers Are Born Biased: Structural Inductive Biases at Random Initialization and Their Practical Consequences**
   Siquan Li*, **Yao Tong***, Haonan Wang*, Tianyang Hu
   Preprints on arXiv, 2026

1. **When Transformers Can (or Can't) Generalize Compositionally? A Data-Distribution Perspective**
   Yao Tong, Jiayuan Ye, Anastasia Borovykh, Reza Shokri
   In NeuriPS Workshop on What Can('t) Transformers Do, 2025

* *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
The Chinese University of Hong Kong 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: ICML 2026, ICLR 2026, NeurIPS 2025 (**Top Reviewer**), ICLR 2025, ICML Workshop 2025, NeurIPS Workshop 2025

Sub-reviewer: CCS 2024, USENIX Security 2024