# Yuetai Li

Second-year Ph. D. student at University of Washington

## EDUCATION

**University of Washington (UW)** Seattle, USA
*Ph. D. student* 2023/09 – Now

- **Advisor:** Professor Radha Poovendran
- **Research Interest:** LLM reasoning, Synthetic dataset, Trustworthy AI

My current research mainly focusses on: (1) understand the reasoning capabilities of LLMs through rigorous analysis (2) investigate synthetic datasets optimized for effective model learning. Previously, I focused on LLM Security and developed defenses against generative backdoor attacks including advertisement injection, code injection, and malicious content generation.

**University of Glasgow (UofG)**, First Honor Degree Glasgow, United Kingdom
*Bachelor of Engineering, Communication Engineering* 2019/09 - 2023/06
**Major GPA:** 3.98/4.00 (94/100), **Ranking:** 2/209

- **Advisor:** Professor Lei Zhang
- **Research Interest:** Distributed Computing, Consensus Algorithm

During my undergraduate studies, my research focused on the theoretical proof and stochastic modeling of distributed consensus algorithms.

- **Scholarships:** James Watt Innovative Talent Scholarship (Top 2%), Academic Scholarship (Top 5%)

## PUBLICATIONS

- Fengqing Jiang, Zhangchen Xu, **Yuetai Li**, Luyao Niu, Bill Yuchen Lin, Radha Poovendran, "SafeChain: Revisiting Safety of Language Model with Long Chain-of-Thought Reasoning Capability", submitted to ACL 2024
- **Yuetai Li**, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Dinuka Sahabandu, Bhaskar Ramasubramanian, Radha Poovendran, "CLEANGEN: Mitigating Backdoor Attacks for Generation Tasks in Large Language Models" Accepted by EMNLP 2024, June 2024
- **Yuetai Li**, Dinuka Sahabandu, Bhaskar Ramasubramanian, Radha Poovendran, "SpecAggre: Spectral Aggregation-based Latent Separability Recovery for Defense against Machine Learning Backdoor Attacks"
- **Yuetai Li**, Zhangchen Xu, Lei Zhang, Jon Crowcroft, "A Modularized Framework of Communication in Consensus", submitted to IEEE/ACM Transactions on Networking
- Zhangchen Xu, **Yuetai Li**, Chenglin Feng, Lei Zhang, "Voting Validity: Exact Fault-tolerant Consensus", Accepted by the 37th IEEE International Parallel & Distributed Processing Symposium (IPDPS), December 2022
- Dachao Yu, Yao Sun, **Yuetai Li**, Lei Zhang and Muhammad Ali Imran, "Communication Resource Allocation of Raft in Wireless Network," Published in IEEE Sensors Journal, September 2023
- **Yuetai Li**, Yixuan Fan, Lei Zhang, Jon Crowcroft, "RAFT Consensus Reliability in Wireless Networks: Probabilistic Analysis", Published in IEEE Internet of Things Journal, February 2023
- **Yuetai Li**, Tong Tong, Benhao Pan, Huajun Yang, Ping Jiang, Weinan Caiyang, "Three-mirror system design for shaping the elliptical beam of a laser diode", Published in Elsevier Optik, Volume 264, 2022

- **Yuetai Li**, Xinbin Chen, Jiale Wang, Tao Zhan, Huajun Yang, Weinan Caiyang, Ping Jiang, "Shaping and transmitting elliptical beam from laser diode by off-axis quadric reflective mirrors", Published in Elsevier Optics Communications, Volume 493, 2021

## SELECTED RESEARCH EXPERIENCES

### Reasoning and Synthetic Dataset

**Small Models Struggle to Learn from Strong Reasoners**

- Show that small models do not consistently benefit from long CoT or distillation from larger models compared to shorter, simpler reasoning chains that better align with their intrinsic learning capacity.
- Proposed Mix Distillation, a simple yet effective strategy that balances reasoning complexity by combining long and short CoT examples or reasoning from both larger and smaller teachers.

**MagpieLM: Synthetic Data Generated from Open-Source LMs.**

- Maintained the official Magpie Hugging Face repository and released open-sourced synthetic data generated from open-source LMs.
- Our aligned MagpieLM models are still SOTA small language models for chat.

SafeChain: Revisiting Safety of Language Model with Long Chain-of-Thought Reasoning Capability

- Investigated how long CoT impacts safety and found that long CoT does not necessarily enhance safety.
- Introduced SafeChain, a dataset designed to improve the safety alignment of LRMs while preserving their reasoning capabilities.

### Trustworthy AI

CLEANGEN: Mitigating Backdoor Attacks for Generation Tasks in Large Language Models

- Proposed CLEANGEN, a novel decoding algorithm that defenses against various backdoor attacks in generation tasks, including **advertisement injection, code injection, and malicious content generation**.
- Proposed theoretical proofs to optimize the decoding overhead and efficiency.

Spectral Aggregation: Latent Separability Recovery for Defense against Backdoor Attacks

- Proposed SpecAggre**,** a novel defense mechanism to aggregate the spectral features of multilayers to detect poisoned samples, restoring the effectiveness of latent separability assumption in adaptive backdoor attacks.

### Distributed Algorithm

Voting Validity: Exact Fault-tolerant Consensus for Preference Aggregation

- Proposed the Voting Validity and the tight bounds of system tolerance to achieve Voting Validity. Designed practical consensus algorithms with proved Termination, Agreement, and Voting Validity.

A Modularized Framework of Communication in Consensus

- Collaborated with Prof. Jon Crowcroft.
- Defined the Reliability Gain and Tolerance Gain formally for the first time, which indicate the logarithmic linear relationship between the consensus reliability and two fundamental network parameters.

## SERVICES

- Teaching Assistant of EE242 Signals and Systems at UW
- Teaching Assistant of EEP595 Network and Communication Security at UW
- Conference Reviewer of ACL Rolling Review (ARR)
- Journal Reviewer of IEEE Internet of Things Journal (IoTJ)
- Journal Reviewer of IEEE Transactions on Network Science and Engineering (TNSE)