

Yanchen Liu

Schwarzman College of Computing
Massachusetts Institute of Technology
Cambridge, MA, USA, 02139

+1 6173970175
ychenliu@mit.edu
liuyanchen1015.github.io

EDUCATION

Massachusetts Institute of Technology (MIT) PhD in Institute for Data, Systems, and Society (IDSS)	2025 - 2029 (expected)
Harvard University MS in Data Science <i>Advisors:</i> Prof. Jiaqi Ma and Prof. Himabindu Lakkaraju	2022 - 2024
Technical University of Munich BS in Computer Science with Highest Honors Minor in Computational Linguistics at Ludwig Maximilian University <i>Advisors:</i> Dr. Timo Schick and Prof. Hinrich Schütze	2018 - 2022

RESEARCH INTERESTS

My research interests mainly lie in **LLM safety**, including robustness [4][5], fairness [3], privacy [1], misinformation [2][6], alignment, jailbreaking [0], unlearning, and, more recently, safety of LM agents [1]. Technically, I have worked on parameter-efficient adaptation/composition [4][5][7] and ICL [3][8].

PUBLICATIONS

- [1] **PrivacyLens: Evaluating Privacy Norm Awareness of Language Models in Action**
Yijia Shao, Tianshi Li, Weiyan Shi, **Yanchen Liu**, Diyi Yang
In Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS 2024)
- [2] **Decoding Susceptibility: Modeling Misbelief to Misinformation Through a Computational Approach**
Yanchen Liu, Mingyu Derek Ma, Wenna Qin, Azure Zhou, Jiaao Chen, Weiyan Shi, Wei Wang, Diyi Yang
In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)
- [3] **Confronting LLMs with Traditional ML: Rethinking the Fairness of Large Language Models in Tabular Classification**
Yanchen Liu, Srishti Gautam, Jiaqi Ma, Himabindu Lakkaraju
In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024)
- [4] **DADA: Dialect Adaptation via Dynamic Aggregation of Linguistic Rules**
Yanchen Liu, William Held, Diyi Yang
In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)
- [5] **Task-Agnostic Low-Rank Adapters for Unseen English Dialects**
Zedian Xiao, William Held, **Yanchen Liu**, Diyi Yang
In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)
- [6] **MIDDAG: Where Does Our News Go? Investigating Information Diffusion via Community-Level Information Pathways**
Mingyu Derek Ma, Alexander K. Taylor, Nuan Wen, **Yanchen Liu**, Po-Nien Kung, Wenna Qin, Shicheng

Wen, Azure Zhou, Diyi Yang, Xuezhe Ma, Nanyun Peng, Wei Wang
In Proceedings of the 38th Annual AAAI Conference on Artificial Intelligence (AAAI 2024 Demonstrations)

[7] [SMoA: Sparse Mixture of Adapters to Mitigate Multiple Dataset Biases](#)

Yanchen Liu*, Jing Yan*, Yan Chen*, Jing Liu, Hua Wu

In ACL Workshop on Trustworthy Natural Language Processing, 2022 (ACLW 2023)

[8] [Semantic-Oriented Unlabeled Priming for Large-Scale Language Models](#)

Yanchen Liu, Timo Schick, Hinrich Schütze

In ACL Workshop on Simple & Efficient Natural Language Processing, 2022 (ACLW 2023)

[9] [Custom Sine Waves Are Enough for Imitation Learning of Bipedal Gaits with Different Styles](#)

Qi Wu, Chong Zhang, **Yanchen Liu**

In Proceedings of the 2022 IEEE International Conference on Mechatronics and Automation (ICMA 2022)

Finalists of Toshio Fukuda Best Paper Award in Mechatronics

RESEARCH EXPERIENCE

Harvard AI4LIFE Group

Research Assistant

Advisor: Prof. [Himabindu Lakkaraju](#)

Mar. 2023 - Present

Cambridge, MA

Stanford NLP Group

Visiting Research Assistant

Advisor: Prof. [Diyi Yang](#)

Oct. 2022 - March. 2024

Palo Alto, CA

LMU Center for Information & Language Processing

Research Assistant

Advisor: Prof. [Hinrich Schütze](#)

Jun. 2021 - Nov. 2021

Munich, DE

ACHIEVEMENTS

[best.in.tum](#)

promotion of the best students

Apr. 2020

TU Munich, DE

TALKS

[Stanford NLP Talk](#)

Dynamic Aggregation and Auto-Discovery of Linguistic Features

Nov. 2023

[Stanford NLP Lightning Talk](#)

LLM for More Research: Empowering Linguistic and CSS Research with LLMs

Oct. 2023

PROFESSIONAL SERVICE

Mentoring:

Rodrigo Nieto - BS/MS@Stanford, Sep. 2023 - Mar. 2024

Mary Williamson - MS@Stanford, Jun. 2023 - Sep. 2023

Reviewer: ACL Rolling Review 2023/2024/2025, COLM 2024/2025, NAACL 2024 SRW, etc.

SKILLS

Programming Languages: C/C++, Java, Python, OCaml, Verilog, MIPS Assembly, SQL...

Language Proficiency: English - TOEFL 111, German - DSH2, Chinese - Native

Also hobbies: Soccer, Go (3 Dan)