Yanchen Liu

Graduate School of Arts and Sciences
Harvard University
Cambridge, MA, USA, 02138

+1 6173970175
yanchenliu@g.harvard.edu
liuyanchen1015.github.io

EDUCATION

Harvard University *2022 - 2024*
MS in Data Science
Cross-Registration in Computer Science at **MIT**

Advisors: Prof. Jiaqi Ma and Prof. Himabindu Lakkaraju

Technical University of Munich *2018 - 2022*
BS in Computer Science with Highest Honors
Minor in Computational Linguistics at **Ludwig Maximilian University**

Advisors: Timo Schick and Prof. Hinrich Schütze
Major GPA: 1.2/1.0 (3.97/4.0) Minor GPA: 1.0/1.0 (4.0/4.0)
Rank: **top 1%** with most courses passed with full scores (1.0/A+), particularly in all math

RESEARCH INTERESTS

My research interests mainly lie in **LLM safety**, including:

1) the **inherent safety of LLMs**: robustness, fairness, privacy, interpretability (overall trustworthiness), and defense against jailbreaking and other malicious activities, etc.; and
2) their safety **when deployed in real-world complex scenarios, interacting with real humans and society**: e.g., as LLM agents, LLM systems, or in social simulations.

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 *Mar. 2023 - Present*
Research Assistant *Cambridge, MA*
Advisor: Prof. Himabindu Lakkaraju

Stanford NLP Group *Oct. 2022 - March. 2024*
Visiting Research Assistant *Palo Alto, CA*
Advisor: Prof. Diyi Yang

LMU Center for Information & Language Processing *Jun. 2021 - Nov. 2021*
Research Assistant *Munich, DE*
Advisor: Prof. Hinrich Schütze

ACHIEVEMENTS

best.in.tum *Apr. 2020*
promotion of the best students *TU Munich, DE*

TALKS

Stanford NLP Talk *Nov. 2023*
Dynamic Aggregation and Auto-Discovery of Linguistic Features

Stanford NLP Lightning Talk *Oct. 2023*
LLM for More Research: Empowering Linguistic and CSS Research with LLMs

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, COLM 2024, NAACL 2024 SRW

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)