

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 - Present

MS in Data Science

Cross-Registration in Computer Science at [MIT](#)

Thesis: Investigating the Fairness of Large Language Models for Predictions on Tabular Data

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](#)

Thesis: Using Unlabeled Examples for Improving Few-Shot Performance of Pre-Trained Language Models

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 lie in **Human-Centered NLP**, with a particular focus on:

- 1) **Learning from Human Language:** understanding, interpreting, and enhancing LLM's behaviors from linguistic perspectives [3][4][7][9];
- 2) **Learning from Human Interaction:** alignment [11], oversight and human-LLM collaboration (e.g. for CSS and linguistic research [9][10]);
- 3) **Ensuring Reliable Human Impact:** reliability, robustness, fairness, and combating misinformation for positive human and social impacts [1][2][3][4][5][6].

PUBLICATIONS

- [1] **From Scroll to Misbelief: Modeling the Unobservable Susceptibility to Misinformation on Social Media**
Yanchen Liu, Mingyu Derek Ma, Wenna Qin, Azure Zhou, Jiaao Chen, Weiyan Shi, Wei Wang, Diyi Yang
Under Review
- [2] **Investigating the Fairness of Large Language Models for Predictions on Tabular Data**
Yanchen Liu, Srishti Gautam, Jiaqi Ma, Himabindu Lakkaraju
Under Review. The Short Version in NeurIPS 2023 Workshop on Socially Responsible Language Modelling Research (NIPSW 2023)
- [3] **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)
- [4] **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)

- [5] [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)
- [6] [SMoA: Sparse Mixture of Adapters to Mitigate Multiple Dataset Biases](#)
Yanchen Liu*, Jing Yan*, Yan Chen*, Jing Liu, Hua Wu
In ACL 2023 Workshop on Trustworthy Natural Language Processing (ACLW 2023)
- [7] [Semantic-Oriented Unlabeled Priming for Large-Scale Language Models.](#)
Yanchen Liu, Timo Schick, Hinrich Schütze
In ACL 2023 Workshop on Simple & Efficient Natural Language Processing (ACLW 2023)
Oral Presentation
- [8] [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

WORKS IN PROGRESS

- [9] [Large Language Models Can Discover Linguistic Features](#)
Yanchen Liu, Mary Williamson, Diyi Yang
- [10] [Let's Do Research Step by Step: Co-Design Your Research Analysis Plan with Large Language Models](#)
Yanchen Liu, Rodrigo Nieto, Diyi Yang
- [11] [Social Gym: Let's Align Step by Step](#)
Ruibo Liu, Jiaao Chen, **Yanchen Liu**, Merrie Morris, Diyi Yang

RESEARCH EXPERIENCE

Stanford NLP Group

Oct. 2022 - Present

Visiting Research Assistant

Palo Alto, CA

Advisor: Prof. [Diyi Yang](#)

- ❑ Developing SocialGym, a simulation sandbox that can serve as both a training and evaluation environment for **social alignment**, enabling LLMs to obtain progressive rewards in simulated social interactions, while also serving as a benchmark to assess LLMs' social alignment in multi-turn dialogues [11].
- ❑ Designing a framework that utilizes LLMs to assist in the step-by-step design of **research analysis plans** for Computational Social Science (CSS) research questions in an interactive and human-AI collaborative manner, exploring how human-AI collaboration can advance social science research [10].
- ❑ Exploring a framework that leverages LLMs to assist humans in verifying and identifying non-standard **linguistic features** in a given text, as well as discovering new linguistic features and usages, demonstrating the potential of empowering linguistic research with LLMs [9].
- ❑ Formulated a computational approach to model users' **susceptibility to misinformation** based on their online activities, using observable sharing behavior as a proxy, and enabling large-scale analysis of its correlation with social and psychological factors [1][5].
- ❑ Proposed **Dialect Adaptation via Dynamic Aggregation (DADA)**, a compositional and modular approach to enhance the dialectal robustness of models trained on Standard American English across multiple dialects simultaneously, from a finer-grained perspective to accommodate dialect flexibility [3].

- ❑ Introduced HyperLoRA, an scalable, task-agnostic method that incorporate expert linguistic knowledge to enable **resource-efficient dialect adaptation** through the use of hypernetworks to disentangle dialect-specific and cross-dialectal information [4].

Harvard AI4LIFE Group

Research Assistant

Advisor: Prof. [Himabindu Lakkaraju](#)

Mar. 2023 - Present

Cambridge, MA

- ❑ Analyzed how LLMs exhibit inherent social biases inherited from their pre-training corpora, and investigated the **fairness implications** of LLMs when making predictions on tabular data, in comparison with traditional machine learning models [2].

LMU Center for Information & Language Processing

Research Assistant

Advisor: Prof. [Hinrich Schütze](#)

Jun. 2021 - Nov. 2021

Munich, DE

- ❑ Proposed **Semantic-Oriented Unlabeled Priming** (Soup), a novel approach by retrieving and leveraging semantically similar unlabeled examples for enhancing the few-shot performance of pre-trained LMs. And introduced bag-of-contexts priming, a new priming strategy that is more suitable for this setting and enables the usage of more examples than fit into the context window [7].

WORK EXPERIENCE

Baidu Inc.

Research Intern

Mar. 2022 - Jul. 2022

Beijing, CN

- ❑ Introduced Sparse Mixture of Adapters (SMoA) to simultaneously mitigate multiple **spurious correlations** in datasets, thereby improving the model's robustness, whereas previous debiasing methods often target a specific bias but fail against others [6].

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

MENTORING

[Rodrigo Nieto](#)

BS, Stanford University

Sep. 2023 - Present

[Azure Zhou](#)

BS, Stanford University

Jun. 2023 - Present

[Mary Williamson](#)

MS, Stanford University

Jun. 2023 - Sep. 2023

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)