FAN YIN

fanyin20@cs.ucla.edu
https://fanyin3639.github.io/

PERSONAL INFORMATION

Gender: Male
Date of Birth (MM/DD/YYYY): 10/16/1997
Place of Birth: Beijing, China
Home Address (US): 1400 Kelton Ave, Unit 202, Los Angeles, CA, 90024
Mobile Phone (US): +1 424 465 0417

EDUCATION

Ph.D student University of California, Los Angeles, CA, USA — September 2020 -
Computer Science department
Graduate research assistant in UCLA-NLP, advisor: Kai-Wei Chang
GPA: 3.83

B.S. Peking University, Beijing, China — September 2016 - July 2020
School of Electronics Engineering and Computer Science
Major: Computer Science
major GPA: 3.62

RESEARCH INTERESTS

Large Language Model, reliability, robustness and safety of NLP systems.

PRE-PRINT

Yihe Deng, Pan Lu, **Fan Yin**, Ziniu Hu, Sheng Shen, James Zou, Kai-Wei Chang, and Wang Wei. Enhancing large vision language models with self-training on image comprehension. In *arxiv*
arxiv

Di Wu, Jia-chen Gu, **Fan Yin**, Nanyun Peng, and Kai-Wei Chang. Synchronous faithfulness monitoring for trustworthy retrieval-augmented generation. In *arxiv*
arxiv

Mohsen Fayyaz, **Fan Yin**, Jiao Sun, and Nanyun Peng. Evaluating human alignment and model faithfulness of llm rationale. In *arxiv*
arxiv

PUBLICATION

Tianyi Yan, Fei Wang, Y James Huang, Wenxuan Zhou, **Fan Yin**, Aram Galstyan, Wenpeng Yin, and Muhao Chen. Contrastive instruction tuning. In *the Annual Meeting of the Association for Computational Linguistics*, 2024
ACL-Findings 2024

Fan Yin, Jayanth Srinivasa, and Kai-Wei Chang. Characterizing truthfulness in large language model generations with local intrinsic dimension. In *International Conference on Machine Learning*, 2024
ICML 2024

Chujie Zheng, **Fan Yin**, Hao Zhou, Fandong Meng, Jie Zhou, Kai-Wei Chang, Minlie Huang, and Peng Nanyun. Prompt-driven llm safeguarding via directed representation optimization. In *International Conference on Machine Learning*, 2024
ICML 2024

Po-Nien Kung, **Fan Yin**, Di Wu, Kai-Wei Chang, and Nanyun Peng. Active instruction tuning: Improving cross-task generalization by training on prompt sensitive tasks. In *Conference on Empirical Methods in Natural Language Processing*, 2023
EMNLP 2023

Da Yin*, Xiao Liu*, **Fan Yin***, Zhong* Ming, Hritik Bansal, Jiawei Han, and Kai-Wei Chang. Dynosaur: A dynamic growth paradigm for instruction-tuning data curation. In *Conference on Empirical Methods in Natural Language Processing*, 2023
EMNLP 2023, equal contribution

Zhouxing Shi*, Yihan Wang*, **Fan Yin***, Xiangning Chen, Kai-Wei Chang, and Cho-Jui Hsieh. Red teaming language model detectors with language models. In *Transactions of the Association for Computational Linguistics*, 2023
TACL, equal contribution with alphabetical order

Fan Yin, Jesse Vig, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. Did you read the instructions? rethinking the effectiveness of task definitions in instruction learning. In *the Annual Meeting of the Association for Computational Linguistics*, 2023
ACL 2023, full paper

Chenghao Yang, **Fan Yin**, He He, Kai-Wei Chang, Xiaofei Ma, and Bing Xiang. Efficient shapley values estimation by amortization for text classification. In *the Annual Meeting of the Association for Computational Linguistics*, 2023
ACL 2023, full paper

Hritik Bansal*, Nishad inghi*, Yu Yang, **FanYin**, Aditya Grover, and Kai-Wei Chang. Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning. In *International Conference on Computer Vision*, 2023
ICCV 2023, full paper

Fan Yin, Yao Li, Cho-Jui Hsieh, and Kai-Wei Chang. Addmu: Detection of far-boundary adversarial examples with data and model uncertainty estimation. In *Conference on Empirical Methods in Natural Language Processing*, 2022
EMNLP 2022, full paper

Fan Yin, Zhouxing Shi, Cho-Jui Hsieh, and Kai-Wei Chang. On the sensitivity and stability of model interpretations. In *the Annual Meeting of the Association for Computational Linguistics*, 2022
ACL 2022, full paper

Fan Yin, Quanyu Long, Tao Meng, and Kai-Wei Chang. On the robustness of language encoders against grammatical errors. In *the Annual Meeting of the Association for Computational Linguistics*, 2020
ACL 2020, full paper

Xiaoya Li*, **Fan Yin***, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. Entity-relation extraction as multi-turn question answering. 2019
ACL 2019, full paper

Yuxian Meng, Wei Wu, Fei Wang, Xiaoya Li, Ping Nie, **Fan Yin**, Muyu Li, Qinghong Han, Xiaofei Sun, and Jiwei Li. Glyce: Glyph-vectors for chinese character representations. 2019
NeurIPS 2019, full paper

WORK EXPERIENCE

Research Intern, Salesforce Research, Palo Alto June 2024 - September 2024
• Work with Philippe Laban, Becky Xiangyu Peng, Yilun Zhou, Jason Wu.

Applied Scientist Intern, Amazon AWS, Santa Clara June 2023 - September 2023
• Work with He He, Samson Tan, Aditya Rawal.

Research Intern, Salesforce Research, Palo Alto June 2022 - September 2022
• Work with Jesse Vig, Philippe Laban, Jason Wu.

Applied Scientist Intern, Amazon AWS, remote June 2021 - September 2021

- Work with He He.

Intern, Beijing ShannonAI Huiyu Technology Co., Ltd January 2019 - December 2019

- Work with Jiwei Li.

TEACHING EXPERIENCE

1. Teaching Associate, UCLA CS M146, Introduction to Machine learning, Fall 2021, with Prof. Kai-Wei Chang
2. Teaching Assistant, UCLA CS M146, Introduction to Machine learning, Fall 2021, with Prof. Kai-Wei Chang
3. Teaching Assistant, UCLA CS M146, Introduction to Machine learning, Winter 2022, with Prof. Sriram Sankararaman
4. Teaching Assistant, UCLA CS M146, Introduction to Machine learning, Spring 2022, with Prof. Aditya Grover

SERVICES

Reviewer: NeurIPS 2024, ICLR 2023, EMNLP 2023, NeurIPS 2022, ARR for ACL 2022, NAACL 2022, NAACL Student Research Workshop 2022.

AWARDS AND HONORS

Excellent Graduate, Peking University July 2020
Merit Student, Peking University October 2019
Peking University Scholarship, Peking University October 2019
2^{nd} Prize in Chinese Physics Olympiad November 2015