Hanqi Yan

PhD Student: hanqi.yan@kcl.ac.uk
University of Warwick $ King's College London (KCL)
Natural Language Processing & Causality in Machine Learning
Homepage: https://hanqi-qi.github.io/

Research Interest

I am a PhD student, doing Natural Language Processing (NLP) and Machine Learning (ML), with a special focus on **interpretable** and **robust** models:

- Considering the stochastic nature of the current deep neural networks, we are able to identify a human-friendly way to understand its model decision-making process.
- The Deep Learning model is built on certain inductive biases and sample selection biases. We propose empirical and principled methods to alleviate representation bias and learn robust representations across various testing environments.
- In the LLM era, I focus on constrained planning (search) in the model decoding phrase, to achieve **safe**, **reliable** and **moral** outputs.

Education

10/2020-	**PhD in Computer Science** University of Warwick & King's College London, United Kingdom Topic: Interpretable and Robust NLP Models Supervisor: Prof. Yulan He
09/2017-07/2020	**Master of Science, Data Science (Computer Science and Technology)** Peking University (PKU), China Topic: Sentiment Analysis and Spatial Data Management Academy for Advanced Interdisciplinary Studies
09/2013-07/2017	**Bachelor of Engineering, Information Technology** Beihang University (BUAA), China Among two students (from department) for advanced entry to PKU

Research Experience

11/2022-02/2023	**Visiting Student in Machine Learning** MBZUAI & CMU Topic: Counterfactual Generation under identifiability Guarantee Advisor: Dr. Kun Zhang (CMU&MBZUAI)
Summer, 2019	**Research Assistant in Computer Science** The Hong Kong Polytechnic University Topic: Causal Reasoning in Sentiment Analysis Advisor: Prof. Wenjie Li

Awards and Honours

2020-2024	The joint scholarship of the China Scholarship Council & University of Warwick
2019	The Research Scholarships at Peking University
2017	Excellent undergraduate thesis at Beihang University

Selected Publications (Core A/A*)

Robust Representation Learning

H. Yan*, L. Kong*, L. Gui, Y. Chi, E. Xing, Y. He, K. Zhang. Counterfactual Generation with identifiability guarantee. ***Neurips23***.

H. Yan*, H. Li*, Y. Li, L. Qian, Y. He and L. Gui. Distinguishability Calibration to In-Context Learning, *Findings of* ***EACL23***.

H. Yan, L. Gui, W. Li, and Y. He. Addressing Token Uniformity in Transformers via Singular Value Transformation, ***UAI22***, ***Spotlight***.

H. Yan, L. Gui, G. Pergola and Y. He. Position Bias Mitigation: A Knowledge-Aware Graph Model for Emotion Cause Extraction, ***ACL21***, ***Oral***.

J. Xu, L. Zhao, **H. Yan**, Q. Zeng, Y. Liang, X. Sun. Lexical-Based Adversarial Reinforcement Training for Robust Sentiment Classification, ***EMNLP19***.

Interpretability based on Generative Model

H. Yan*, L. Gui*, and Y. He. Hierarchical Interpretation of Neural Text Classification, *Computational Linguistics, presented in* ***EMNLP22***.

H. Yan, L. Gui, M. Wang, K. Zhang and Y. He. Explainable Recommender with Geometric Information Bottleneck. Under Review of **TKDE**.

Large Language Model

Y. Zhou, J. Li, Y.Xiang, **H.Yan**, L. Gui, Y. He. The Mystery and Fascination of LLMs: A Comprehensive Survey on the Interpretation and Analysis of Emergent Abilities. Under Review of **TACL**.

H. Yan, Q. Zhu, X. Wang, L. Gui, Y. He. Steer the LLMs in the self-refinement loop via unsupervised Reward. Under Review of **ARR**.

Professional Activities

Event Organizer

- Co-Chair of AACL-IJCNLP (Student Research Workshop), 2022.

Reviewer

- ACL23', EMNLP22'23', NAACL24', EACL23', AACL24', UAI23', AISTATS24'; Neurocomputing, Transactions on Information Systems (TOIS)

Conference Oral Presenter

- ACL21', Oral, Remote.
- UAI22', Spotlight, Eindhoven.

Conference Poster Presenter

- UAI22', spotlight, Eindhoven.
- EMNLP23', Abu Dhabi.
- ICML23', Counterfactuals in Minds and Machines Workshop, Hawaii.
- Neurips23', Poster, New Orleans.

Course Teaching

University of Warwick, Natural Language Processing. 2021,2024 Spring/2023 Fall.
University of Warwick, Web Development Technologies, 2021 Fall.
Peking University, Teaching assistant of Introduction to Aerospace Engineering, 2018 Fall.

Language

Chinese (Native speaker), English (Working Proficiency), Cantonese (Basic)