Hanqi Yan

Postdoctoral researcher: hanqi.yan@kcl.ac.uk
King's College London (KCL)
Natural Language Processing & Causality in Machine Learning
Homepage: https://hanqi-qi.github.io/homepage/

Research Interest

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

- Conceptualize the latent variables and involve them into the model decision-making process to establish faithful *self-explanatory* models.
- Empirical and principled methods to enhance model robustness over various test inputs, especially on *distribution shifts* and *structural limitations in Transformers*.
- In the Large Language Model era, I focus on (a) controllable generation to achieve *safe*, *reliable* and *moral* outputs; (b) reasoning abilities enhancement via *planning*; (c) representation learning to address the vulnerability to trivial input perturbations.

Education

10/2020-04/2024	**PhD in Computer Science** University of Warwick, 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

01/2024-	**PostDoc in Informatics** King's College London (KCL) Topic: Robust and Reliable Language Models
11/2022-02/2023	**Visiting Student in Machine Learning** MBZUAI & CMU Topic: Counterfactual Generation under identifiability Guarantee Advisor: Prof. Kun Zhang (CMU&MBZUAI)
Summer, 2019	**Research Assistant in Computing** The Hong Kong Polytechnic University Topic: Causal Reasoning in Sentiment Analysis Advisor: Prof. Wenjie Li

Awards and Honours

04/2024	PhD viva with no Corrections. Examiners: Prof. Theo Damoulas, Prof. Yonatan Belinkov
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 and Manuscripts (Core A/A*)

Large Language Model

H. Yan, Q. Zhu, X. Wang, L. Gui, Y. He. Mirror: A Multiple-perspective Self-Reflection Method for Knowledge-rich Reasoning. **ACL24**.

Y.Xiang, **H.Yan**, L.Gui, Y. He. Addressing Order Sensitivity of In-Context Demonstration Examples in Causal Language Models. **ACL24-findings**.

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.

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

Robust Representation Learning

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

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. **TKDE**.

Other Topics

Y Lei, H Pei, **H Yan**, W Li. Reinforcement learning based recommendation with graph convolutional q-network. ***SIGIR2020***

R Zhao, L Gui, **H Yan**, Y He. Tracking Brand-Associated Polarity-Bearing Topics in User Reviews. *TACL2023*, presented at ***ACL23***

Professional Activities

Event Organizer:

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

Reviewer:

NLP: ACL23'24', EMNLP22'23'24', NAACL24', EACL23', AACL24'
Machine Learning: UAI23', AISTATS24', Neurocomputing, Transactions on Information Systems

Conference Oral Presenter:

NLP: ACL21'(Remote)
Machine Learning: UAI22' (Eindhoven)

Conference Poster Presenter:

NLP: EMNLP23'(Abu Dhabi),
Machine Learning: UAI22'(Eindhoven), ICML23'(Hawaii), Neurips23'(New Orleans)

Invited Talk

UC San Diego, invited by Prof. Zhiting Hu. 02/2024

- Title: *Robust and Interpretable NLP via representation learning and Path Ahead*

Yale University, invited by Prof. Arman Cohan. 01/2024

- Title: *Robust and Interpretable NLP via representation learning and Path Ahead*

Turing AI Fellowship Event, London, 03/2023

-Title: *Distinguishability Calibration to In-Context Learning*

UKRI Fellows Workshop, University of Edinburgh, 04/2022.

- Title: *Interpreting Long Documents and Recommendation Systems via Latent Variable Models*

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

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