Hanqi Yan

PhD Student	Email: Hanqi.Yan@warwick.ac.uk
Computer Science	Google Scholar
University of Warwick, UK	Homepage

Research Interest

I am a PhD student, doing Natural Language Processing (NLP) and Machine Learning, with a special focus on interpretable and robust models. Recently, I primarily focus on the two directions:

- Considering the stochastic nature of the current deep neural networks, we are able to identify a human-friendly way to present its learning process by Bayesian probabilistic explanations.
- The Deep Learning model is inevitably to be biased and vulnerable, as it is built on certain inductive biases and sample selection bias. We could inject the human-readable prior knowledge to calibrate the model output and build a robust model.

Education

10/2020-10/2024 (expected)	**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, China Topic: Sentiment Analysis and Spatial Data Management
09/2013-07/2017	**Bachelor of Engineering, Information Technology** Beihang University (BUAA), China

Research Experience

03/2023-	**Visiting Student at the Department of Informatics** Kings' College London (KCL), United Kingdom Topic: Trustworthy generative AI Advisor: Prof. Yulan He (Warwick&KCL)
11/2022-02/2023	**Visiting Student in Machine Learning** MBZUAI, United Arab Emirates 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*)

Large-pretrained model calibration

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*.

Robustness of sentiment analysis model

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.
H. Yan*, L. Kong*, L. Gui, Y. Chi, E. Xing, Y. He, K. Zhang. Counterfactual Generation under identifiability guarantee. ICML23, workshop.

Professional Activities

Event Organizer

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

Reviewer

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

Conference Oral Presenter

- Annual Meeting of the Association for Computational Linguistic (ACL21', oral), Remote.
- Conference on Uncertainty in Artificial Intelligence (UAI22', spotlight), Eindhoven.

Conference Poster Presenter

- Conference on Uncertainty in Artificial Intelligence (UAI22'), Eindhoven.
- Conference on Empirical Methods in NLP (EMNLP23'), Abu Dhabi.
- Conference on Uncertainty in Artificial Intelligence (ICML23', Workshop), Hawaii.

Course Teaching

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

Skills

Programming: Python, PyTorch
Deep Learning Framework: Transformers, Variational AutoEncoder, Adversarial Training
Others: Prompt Engineering, Knowledge Graph

Language

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