Research Statement

Causal Artificial Intelligence and Data Science

Linyi Yang, Ph.D.
yanglinyi@westlake.edu.cn

1 Research Overview:

– Correlation $\neq$ Causation

Research in Artificial Intelligence (AI) has advanced at an incredible pace, to the point where it is making its way into our everyday lives, explicitly and behind the scenes. In particular, the recent rapid progress in large language models (LLMs) has set off a new wave of AI enthusiasm. However, beneath their impressive progress, many AI models suffer from a lack of factuality control, weakness in out-of-distribution (OOD) generalization, hallucination, and other issues, the cause of which can be attributed to the underlying mechanisms of training a deep learning model by minimizing empirical risks over training data, not capturing core causal features. The presence of these issues raises a question at the heart of my research focus: **How do we make the theoretical and empirical combination of recent advances in causal reasoning techniques and deep learning models?**

I strive to empower deep-learning models with causal-thinking methods and pay particular attention to data-driven fields, aiming to improve data efficiency, robustness, interpretability, and OOD generalization of AI systems. On the one hand, I design ★1) **unified out-of-distribution benchmarks and interpretable analyses using practical settings to identify the robustness and stability issues of current methods**. The pursuit of interpretability functions as a metaphorical X-ray, shedding light on the internal workings of black-box models, thereby enabling both researchers and industrial practitioners to gain a clear understanding of predictions. On the other hand, despite experts' best efforts, current LLMs are almost guaranteed to be imperfect due to disparities between training methodologies (empirical risk minimization, ERM) and the practical deployment landscapes, leading to the pervasive "hallucination" issue. To make AIs more effective for downstream applications, ★2) **I also develop data-centric algorithms for mitigating spurious patterns learned by models and enhancing their OOD robustness.** In addition, I realize the ongoing significance of robustness and hallucination issues in the era of large language models [2, 15]. It's crucial to understand how to make robust and generalizable causal and counterfactual statements in the context of heterogeneous and biased data collections, including confounding bias, selection bias, dataset shift, and issues of transportability.

My research probes the intersection between data mining and natural language processing (NLP). It has led to publications in top-tier conferences and journals in both areas (*e.g.,* WWW, CIKM, ACL, AAAI, NeurIPS) [11, 13, 7, 14, 4]. I conduct *post-hoc analysis* to identify pitfalls in current natural language understanding systems using counterfactual explanations [10, 7], and design unified OOD benchmarks [9] as well as causality-inspired evaluation tools [5]. In addition to identifying research problems, I also try to solve them with data-driven approaches [6, 8, 12], inspired by human rationales using automatic semi-factual data augmentation. The impact of my work extends beyond academics: several of the frameworks I developed have been integrated into open-source AI libraries or deployed internally in industries. For instance, one of the projects I was involved in has been downloaded more than 48,000 times on Huggingface per month [3]. In the future, I am eager to continue improving AL models in *high-stakes, in-the-wild* scenarios using causal-thinking methods. I also plan to help models better deal with biased data collection, which is crucial for pushing the limits of model usability. I will continue to secure more external research funding through the research council, the Ministry of Science and Technology, and industries. I will keep a high publication rate and make attempts to produce quality and influential outputs.

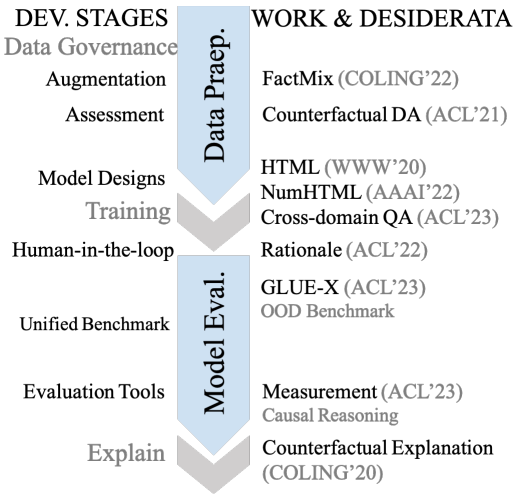

Figure 1: I support experts throughout the AI development cycle. Given a development stage (left), I distil its unique desiderata and design NLP models and evaluations accordingly (right).

2 Future Research Agenda:

My long-term research goal is to reduce the likelihood of hallucinations in AI models to the same level as in humans under the theory of the Structural Causal Model (SCM), which provides a coherent mathematical foundation for the analysis of causes and counterfactuals. In my past work, I achieved interpretable analysis and data-efficient methods primarily for pre-trained language models by explaining counterfactuals and semi-factuals. Nevertheless, the emergence of LLMs brings opportunities that can support a wider variety of applications and people involved. Meanwhile, new research questions arise in the academic field. How do we reduce the hallucinations in LLMs? How do we deal with label inconsistency for naturally subjective applications? How do we prevent humans from being misled by unfaithful responses generated by LLMs? In the coming years, I intend to answer these questions by **enhancing LLMs' awareness of (1) causal-thinking capabilities, and (2) faithfulness**.

To make model training and analysis more human-like and reflective of our complex real-world goals, I will work with domain experts to explore more efficient ways to collect and use benchmark datasets in the context of high-stakes applications (e.g., in medicine, biology, and science). It is anticipated that this work will therefore benefit China's and Westlake's strategic areas of causal AI, data science, interpretable ML [1], and AI4Science [16]. To carry out the research in a systematic way, I divide the research activities at Westlake University into 1 managerial and 3 technical work packages (WPs), with a focus on more thoughtful data governance and model training.

The management work package is run by the Westlake Fellow with day-to-day assistance from RA or Postdoc researchers. The internal coherence of the work and existing research achievements should help ensure good management of the project, including making sure that the project is producing high-quality research of international excellence, strong engagement with the industrial partners and actively seeking financial contributions. Avail appropriate opportunities for dissemination and outreach and engage with relevant groups, and provide an easy-to-understand brochure of the project's aims and rationale for public distribution.

WP1 – Causal Explanations in AI4Science (Start-up Fund)

I expand my research scope to include a wider of people and look forward to collaborating with experts in other domains. I also believe most insights from my work are transferable to AI applications beyond NLP, including:

1. Learn causal invariance for OOD generalization on molecule representation learning and drug discovery tasks;
2. Generate causal explanations for any graph neural networks (GNNs) based on learned latent causal factors;
3. Deploy causality-inspired explainable systems in the field of AI4Science, such as computational biology.

WP2 – Causality-Aware Data Governance and Model Training (Co-PI of National Key Project)

I plan to leverage causal principles to design more robust, interpretable, and controllable models as follows:

1. Distill unique requirements for each task and tailor tool designs to mitigate biased data collections;
2. Enhance the value of counterfactuals for training robust models, by adding explicit terms in the loss function that compare counterfactuals with original data;
3. Construct trustworthy benchmarks by using causality-inspired evaluation metrics.

WP3 – Reduce the Likelihood of Hallucinations in LLMs (Apply to NSFC Fund)

Through longitudinal studies, I strive to make LLMs more generalizable, interpretable, and faithful by using supervised knowledge from fine-tuned models or external databases, including:

1. Design automatic hallucination detection tools for revealing the faithfulness issues of current LLMs;
2. Enhance the fact-checking ability of LLMs by attaching an external database;
3. Use PLMs to improve the faithfulness of LLMs in natural language understanding and generation tasks.

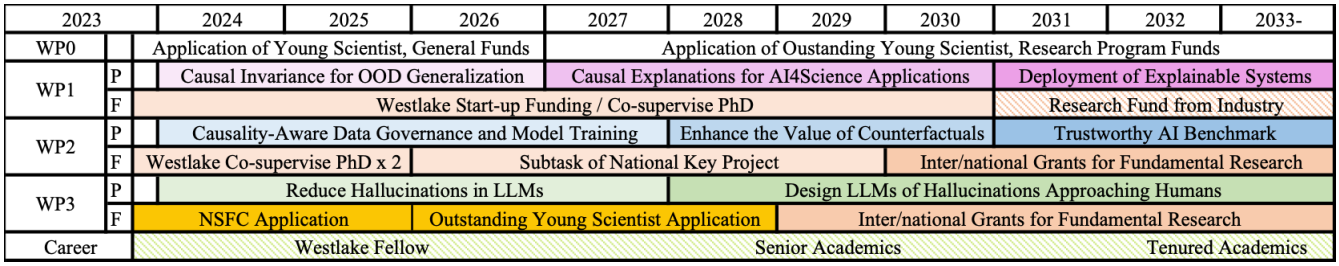

Figure 2: The plan for the applicant's academic career development. WP: Work Package, P: Project, F: Fund.

References

[1] Valerie Chen et al. "Best practices for interpretable machine learning in computational biology". In: *bioRxiv* (2022), pp. 2022–10.

[2] Jindong Wang et al. "On the robustness of chatgpt: An adversarial and out-of-distribution perspective". In: *arXiv preprint arXiv:2302.12095* (2023).

[3] Yidong Wang et al. "PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization". In: *arXiv preprint arXiv:2306.05087* (2023).

[4] Yidong Wang et al. "Usb: A unified semi-supervised learning benchmark for classification". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 3938–3961.

[5] Linyi Yang, Yingpeng Ma, and Yue Zhang. "Measuring Consistency in Text-based Financial Forecasting Models". In: *arXiv preprint arXiv:2305.08524* (2023).

[6] Linyi Yang et al. "A Rationale-Centric Framework for Human-in-the-loop Machine Learning". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2022.

[7] Linyi Yang et al. "Exploring the Efficacy of Automatically Generated Counterfactuals for Sentiment Analysis". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021, pp. 306–316.

[8] Linyi Yang et al. "FactMix: Using a Few Labeled In-domain Examples to Generalize to Cross-domain Named Entity Recognition". In: *Proceedings of the 29th International Conference on Computational Linguistics*. 2022, pp. 5360–5371.

[9] Linyi Yang et al. "GLUE-X: Evaluating Natural Language Understanding Models from an Out-of-distribution Generalization Perspective". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023.

[10] Linyi Yang et al. "Generating Plausible Counterfactual Explanations for Deep Transformers in Financial Text Classification". In: *Proceedings of the 28th International Conference on Computational Linguistics*. 2020, pp. 6150–6160.

[11] Linyi Yang et al. "Html: Hierarchical transformer-based multi-task learning for volatility prediction". In: *Proceedings of The Web Conference 2020*. 2020, pp. 441–451.

[12] Linyi Yang et al. "Learning to Generalize for Cross-domain QA". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023.

[13] Linyi Yang* et al. "Maec: A multimodal aligned earnings conference call dataset for financial risk prediction". In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2020, pp. 3063–3070.

[14] Linyi Yang et al. "NumHTML: Numeric-Oriented Hierarchical Transformer Model for Multi-task Financial Forecasting". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 10. 2022, pp. 11604–11612.

[15] Linyi Yang et al. "Out-of-Distribution Generalization in Text Classification: Past, Present, and Future". In: *arXiv preprint arXiv:2305.14104* (2023).

[16] Nianzu Yang et al. "Learning substructure invariance for out-of-distribution molecular representations". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 12964–12978.