Tianyi(Lorena) Yan tianyiy@usc.edu

As the landscape of natural language processing (NLP) evolves, we are ushered into an era where language models (LMs) with billions of parameters become widely used. While these models have achieved impressive feats on various benchmarks, they can still lack robustness to unseen data and fail to produce desired outputs. In addition, the low interpretability of black-box models causes difficulties in systematically detecting confounders and enhancing their robustness. These issues motivate me to place my research interests in the following two questions: **(1) how to provide more targeted guidance to make LM more controllable and reliable, and (2) how to further enhance LMs' consistency and robustness with better interpretability and evaluation benchmarks.**

More targeted guidance for developing controllable and reliable LMs My research journey began with exploring LMs' robustness to unintended dataset biases under the guidance of Prof. Muhao Chen, culminating in our ACL 2023 publication [1]. Initially, I conducted preliminary experiments that combined predictions from pretrained teacher and main models but gained limited improvement. Because low-level layers can preserve rich surface features [2], we hypothesized that the main model can still suffer from biased attention patterns that overly relies on shortcuts like lexical overlap in natural language inference. Therefore, adding to the Product-of-Expert method that multiplies top-level predictions from a biased model and our main model, we also combine lower-level attention distribution additively. This approach significantly enhanced the model's out-of-distribution (OOD) performance by encouraging the biased model to capture spurious shortcuts and allowing the main model to fit the unbiased attention pattern. I further analyzed the attention distribution and discovered that our method successfully directed more attention to non-overlapping tokens that are more important for paraphrase identification (one of the tasks we focused on).

Motivated by the intriguing findings, I started to ask: how can we further harness attention or other signals to encourage the model to be more **controllable and reliable**? Seeing that an effective debiasing method may direct the model to rely less on shortcuts, I am eager to explore how we can further guide LMs to place less attention on undesired features and **unlearn misalignments** such as stereotypical biases and toxicity or place more attention that end-user wants to emphasize during inference time. For instance, one interesting approach I want to explore is leveraging reinforcement learning (RL) and the idea of conditional transformer [3] to condition LMs' attention distribution on special control tokens. I want to train the model to associate these tokens and align attention with its output sequences with higher rewards if they adhere to specific preferences, such as reducing repetitions or minimizing toxicity. Delving into this direction, I want to develop generalizable methods that help LMs steer themselves toward desired outputs and become more **responsible and trustworthy**.

Enhancing consistency and generalizability with better interpretability and benchmark In another paper under submission [5], I furthered my interest in LMs' **robustness and consistency**. While these models can generate coherent texts, they still exhibit high variance in performance given the same instruction with different forms and formulations. To address this issue, I implemented a contrastive learning approach that increases the similarity between hidden representations of instructions that are semantically equivalent. Experimenting with different learning objectives at instruction and instance levels and varying selections of hard examples, my method enhanced LM's consistency more effectively when contrastive learning is performed on the last token of the output sequence for generative tasks and the class label for discriminative tasks.

With the positive results, new questions came to my mind: when I am aligning representations of semantically similar instructions of different tasks, how am I reshaping its latent space? Is it possible to **localize components** of an LM that are more responsible for capturing task-specific versus cross-task knowledge? Surprised by how LLMs can excel in generative tasks but can still make basic errors in the discriminative version of the same task [4], which are usually assumed to be easier, I want to develop **more interpretable models by attributing knowledge to specific subnetworks**. I believe we can then edit LMs more systematically, mitigate the discrepancy between generating and understanding, and enhance their invariant knowledge base as a stronger foundation for bolstering robustness and consistency.

In addition, I noticed a lack of reliable benchmarks for evaluating LMs' OOD performance given the trend of automatic instruction data generation, as it is difficult to guarantee the test set is not contaminated. Given this situation, I leveraged sentence transformers to measure the semantic similarity between the training and test data for measuring LLMs' performance on unseen data. Nonetheless, stricter evaluation benchmarks are needed to evaluate LMs' robustness and generalizability. Thus, I am also interested in exploring **provable guarantees for dataset contaminations and OOD detection.**

Career Path In the long term, I want to become a professor. Working as a teaching assistant for introduction to machine learning and algorithm courses at USC, I enjoy the rewarding experience of guiding students and fostering a collaborative and supportive learning environment. Also, after benefitting from great mentorship from my Ph.D. mentors and great faculty advisors, I want to pay back to academia by passing down knowledge to researchers of future generations.

References

[1] Fei Wang*, James Y. Huang*, **Tianyi Yan**, Wenxuan Zhou, and Muhao Chen. Robust Natural Language Understanding with Residual Attention Debiasing. In *Findings of the Association for Computational Linguistics,* 2023.

[2] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, 2019.

[3] Nitish Shirish Keskar*, Bryan McCann*, Lav R. Varshney, Caiming Xiong, and Richard Socher. CTRL: A Conditional Transformer Language Model For Controllable Generation, *arXiv:1909.05858*, 2019.

[4] Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D. Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Raghavi Chandu, Benjamin Newman, Pang Wei Koh, Allyson Ettinger, and Yejin Choi. The Generative AI Paradox: "What It Can Create, It May Not Understand". *arXiv preprint arXiv:2311.00059,* 2023.

[5] **Tianyi Yan**, Fei Wang, James Yipeng Huang, Wenxuan Zhou, Wenpeng Yin, and Muhao Chen. Contrastive Instruction Tuning (Paper under submission)