My research interests are in Machine Learning (ML) and Natural Language Processing (NLP), particularly in its intersections with Robustness and Generalization. Current machine learning techniques fall short of human abilities in both their capacity to learn with causal reasoning and to learn robust prediction mechanisms. My research aims to (1) improve pre-trained models and distill knowledge from large corpus, and (2) understand and improve the robustness in learning. I have been working with several amazing professors across different institutions in the United States since my first year at the University of Minnesota, Twin Cities, and I have been fortunate to conduct some initial works aligned with my research goals underpinned by advanced knowledge well absorbed from intense graduate courses I had with excellent grades as well as essential mathematical ability I mastered through them. To further fulfill my goals, I would like to apply for your prestigious PhD program.

Improve Pre-trained Models and Distill Knowledge from Large Corpus How to distill the knowledge from large corpus poses a crucial problem for large pre-training models. Reining in our models with appropriate human priority, such as causality and linguistic attention, will aid generalization across inputs and may also provide the possibility for transferring the knowledge from large corpus into specific domain desired.

My impetus for pursuing this research direction comes from experiences during a research internship at *the CLIP Laboratory at* ***University of Maryland***, where I worked with ***Professor Jordan Boyd-Graber*** on improving Open-domain Question Answering systems by transferring knowledge from a new, out-of-domain question answering dataset (QuizBowl). Approaches to answer QuizBowl questions directly are complicated because they must decode the convoluted syntax, select useful information that cannot be transformed directly through different formats. In the sequential papers, ***Improving Question Answering with Generation of NQ-like Questions*** submitted to MRQA 2021 and ***You Make me Feel like a Natural Question: Training QA Systems on Transformed Trivia Questions*** submitted to EMNLP 2022, I co-led a project to avoid such baroque complexity and to use human priority to select the pieces of information. We proposed systems to automatically generate shorter, information-seeking questions, resembling web queries in the style of Natural Questions (NQ) dataset from longer trivia questions. We boosted the pre-trained retriever to better search correct sets of contexts by leveraging the page information from the QuizBowl metadata into question generation. Our system fine-grained study what specifically attributes to better question generation for the domain desired and further fine-tune the Dense Passage Retriever for better domain adaptation.

I am always interested in exploring how causality can help us improve the robustness in neural systems and designing systems with a focus on causal features even under environmental variations, a research direction sparked during my time in the ***UMN NLP group*** under ***Professor Dongyeop Kang***.

In a paper submitted to EACL 2023, ***Debiasing Language Models for In-Context Learning by Adapting Causal Inference***, I led a project about de-biasing pre-trained language models (PLMs) for in-context learning. I noted that common prompting practice for PLMs rather elicits a summary of observational data (i.e. pre-training corpora) than measuring the true causal effect of the input text on possible labels. And our work showed that under standard prompting practice, confounding shifts can actually deteriorate the in-context learning accuracy of PLMs. We've mitigated such confounding bias and improved model accuracy by methods inspired from causal inference, substantially increasing the accuracy of PLMs in three classification tasks and

reducing accuracy variance across different class distributions in prompts to achieve better robustness.

To Understand and Improve the Robustness in Learning Under the torrent of large pre-trained language models paradigm in the context of NLP, I want to research on the relations between overparameterization, generalization with robustness in neural systems. Through a summer research internship at ***University of Pennsylvania*** with ***Professor Weijie Su***, I lead a project about a new theory of adversarial examples and generalization in neural networks. We analyze the reasons for the existence of adversarial examples based on the perspective of the length of norms from perturbations and define different types of adversarial examples based on the elasticity property of target labels during neutral training. Our work helps to understand the relations between robustness and generalization.

Diffusion Models provide me with a new perspective for robust learning – adversarial purification. Traditional adversarial training, which trains neural networks on adversarial examples, can only defend against a specific attack that they are trained with high computational complexity, thus leaving a huge potential for improvements. Adversarial purification, relying on generative models to purify adversarially perturbed samples before classification, has emerged to defend against unseen threats in a plug-n-play manner without re-training the classifiers. My impetus for pursuing this research direction comes from the recent collaboration with ***SONY*** within the ***UMN NLP group.*** We work on better decoders for textual generative models. When I adapted Diffusion-LM for more composable text operations, specifically to edit the text w.r.t. an attribute and to manipulate keywords, the controllable generation ability and strong sample quality from Diffusion Models amazed me and helped me to think of it as an ideal candidate for contextual adversarial purification. I'm adapting Diffusion Models for out-of-distribution robustness which concerns adversarial robustness under domain shifts. I'm utilizing the controllable generation ability as a preprocessor for zero-shot domain adaptation (e.g. the same parse tree or POS tagging as the target domain paradigm), while the strong sample quality for purifying adversarial texts. Our work also analyzes the possibility of augmenting intermediate diffusion samples for robust training.

Future Plans My career aspiration is to become a professor, since an academic career offers unique opportunities to detect the impactable research questions and to solve them, while also to mentor students through research advising. This choice is particularly informed by my positive experiences as an undergraduate research assistant. I have fully enjoyed doing research and advising junior members in the group!

Why School Paragraph