My research interests are in Machine Learning (ML) and Natural Language Processing (NLP), more specifically how they intersect with Robustness and Generalization. Current machine learning techniques fall short of human abilities in both their capacity to learn with causal reasoning and to develop robust prediction mechanisms. My research aims to (1) improve pre-trained models and distill knowledge from large datasets, and (2) understand and improve the robustness in learning.

I have been working with several leading professors across different institutions nationwide since my first year at the University of Minnesota, Twin Cities. I have also been fortunate to conduct some initial works aligned with my research goals underpinned by my advanced knowledge well absorbed from intense graduate courses, resulting in excellent grades, as well as essential mathematical abilities I mastered through them. Given my interests and pursuits, I believe your PhD program is the ideal next step in my journey.

Improve Pre-trained Models and Distill Knowledge from Large Corpus

My impetus for pursuing this research direction came from experiences during my research internship at *the CLIP Laboratory at the* ***University of Maryland***. 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 are complicated because they must decode the convoluted syntax through selecting useful information that cannot be transformed directly within multiple 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 use human priorities to select the relevant pieces of information. I contributed to proposing systems that would automatically generate shorter, information-seeking questions, resembling web queries in the style of Natural Questions (NQ) dataset from longer trivia questions. My group and I improved the performance of open-domain question answering under domain shifts by leveraging the page information from metadata into question generation to boost the pre-trained retriever, and further fine-tuning the Dense Passage Retriever for better domain adaptation. During this experience, I found that reining in models with appropriate human priorities aids generalization across inputs and may also allow the transfer of knowledge from the large corpus into a specific domain.

Causality, as another human priority, inspired me to explore how it can help improve the robustness of neural systems and designing systems with a focus on causal features even under environmental variations. This research direction sparked during my time in the ***UMN NLP group*** under ***Professor Dongyeop Kang***.

In leading a project about de-biasing pre-trained language models (PLMs) for in-context learning, I noted that common prompting practice for PLMs elicits a summary of observational data (i.e. pre-training corpora) rather than measuring the true causal effect of the input text on possible labels. Our work showed that under standard prompting practice, confounding shifts can actually deteriorate the in-context learning accuracy of PLMs. We mitigated such confounding biases 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. This paper, ***Debiasing Language Models for In-Context Learning by Adapting Causal Inference***, has since been submitted to EACL 2023. This work developed my ambitions to bring causal reasoning into large language models.

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 the relationship between overparameterization and generalization with robustness in neural systems. Through a summer research internship at the ***University of Pennsylvania*** with ***Professor Weijie Su***, I led a project assessing the theory of adversarial examples and generalization in neural networks. We analyzed the existence of adversarial examples from the perspective of perturbation norm lengths and defined different types of adversarial examples based on the elasticity property of target labels during neutral training. This work provided me with a basic sense for the relations between robustness and generalization, which inspired me to further improve robust learning.

Diffusion Models have given me a new perspective on robust learning, introducing me to adversarial purification. Traditional adversarial training, which trains neural networks on adversarial examples, can only defend against specific attacks that they are trained on, thus leaving tremendous potential for improvements. Adversarial purification, relying on generative models to purify adversarially perturbed samples before classification, has emerged as a way to defend against unseen threats in a plug-n-play manner without re-training the classifiers. My desire to pursue this research direction comes from my recent collaboration with ***SONY*** within the ***UMN NLP group,*** where I have worked on creating better decoders for textual generative models. When I adapted Diffusion-LM to facilitate more composable textual operations, specifically to edit the text with regards to an attribute and to manipulate keywords, the controllable generation ability and strong sample quality of Diffusion Models amazed me and inspired me to think of it as an excellent candidate to assess out-of-distribution robustness, which concerns adversarial robustness in the context of domain shifts. I am 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 to purify adversarial texts. Our work also studies the possibility of augmenting intermediate diffusion samples for robust training. Our work aims to submit to the ACL 2023 conference.

My career aspiration is to become a professor given that academic careers offer unique opportunities to be involved in cutting-edge research, solving complex problems, and mentoring budding researchers and students. This direction has been informed especially by my positive experiences as an undergraduate research assistant as I have thoroughly enjoyed conducting research and advising junior members in the group.

At JHU, I am especially interested in being co-advised by ***Professor Jason Eisner*** and ***Professor Daniel Khashabi*** to continue my work at the intersection of NLP and foundational deep learning. I hope to work with ***Professor Jason Eisner*** on better decoders for generative models and diffusion models for compositional controllable text generations, and critical analysis for autoregressive models and diffusion models. I hope to work with ***Professor Daniel Khashabi*** on representation learning for long sequences combining spatiality of transformer and temporality of diffusion models, and also on convergence analysis in KL Divergence for diffusion models and better sampling algorithms for score-based generative models. I am also interested in working with ***Professor Anjalie Field*** on social-aware language modeling and bias mitigation in PLMs. It would also be constructive to work with ***Professor Anqi Liu*** on robust ML, studying OOD problems via local manifold smoothness and local elasticity. Following the work of the Center for Language and Speech Processing (CLSP) group has led me to see a clear fit for my skills and interests at JHU, and I am confident that it is the ideal place for me to pursue my PhD.