My research interests lie in Natural Language Processing (NLP) with a focus on **Trustful NLP** and **Compositional and Grounded NLP**. I have been fortunate to embark on a research journey towards these directions with my advisors Professors Christopher Potts and Desmond Ong. These initial steps along with career plans are discussed below.

Trustful NLP A trustful NLP system needs to be interpretable, robust, and life-long. Currently, a neural model often lacks such capabilities, as it works as a black-box that leverages statistical biases and artifacts to achieve unsystematic solutions.

Humans have well-founded knowledge about causal structure, ranging from commonsense intuitions to advanced scientific knowledge. I am interested in realizing these symbolic insights in a neural model while allowing it to be trained in a data-driven fashion. Specifically, a neural model that aligns with targeted symbolic computations can perform modular computations and preserve information locality. Consequently, the neural model becomes more systematic and interpretable. In parallel, I am also interested in eliminating these barriers by using data-centric approaches, such as making new datasets or augmenting existing datasets to train neural models to be more trustful.

My motivation for pursuing this research direction is from interpreting self-attention models for sentiment analysis. After working on a Transformer-based model for sentiment analysis task [1], we were interested in interpreting the learned attention weights. In our paper presented at EMNLP 2020 Workshop on BlackboxNLP [2], we developed our attention attribution method to aggregate self-attention through layers, and found that self-attention weights encode task-specific semantics. We extended this work by formally comparing our method with other gradient-based attribution methods on sequence classification tasks [3]. In our recent paper presented at AAAI 2021 [4], we went forward to enhance the formulation of self-attention weights by bringing in negative weights. We showed that such richer representations of attention weights improved performance as well as interpretability. I would like to continue this journey by studying the mechanism of self-attention weight that seems to serve as an information bottleneck that compresses information in the way that is most relevant to a task.

My firsthand experiences with robustness in NLP come from investigating model performance on adversarially created datasets. My co-first authored paper presented at ACL 2021 [5] showed neural models could be fooled with hand-crafted sentences for sentiment analysis. With successive rounds of data collection and model training, we found that neural models transfer knowledge less effectively when newer rounds containing large label distribution shift. To further investigate the transfer learning mechanism, I am working to identify the limits of cross-task transfer under extreme conditions, such as knowledge transfer without word identities [6]. Our results suggest that models may learn non-linguistic knowledge such as statistical priors including dimension reduction and clustering. I plan to further investigate a more efficient pre-training pipeline that induces non-linguistic priors. Additionally, I would be keen to study how to fine-tune models under a known label distribution shift to achieve life-long learning with dynamic benchmarks.

I also am co-leading an ongoing project towards building neural models that conform to targeted causal structures. Through additional training signals to guide neural models to realize pre-defined causal structures, we demonstrate that it substantially improves the model's interpretability and generalizability [7]. In a similar vein, I am leading a project showing that causal abstraction brings performance gains over alternatives like model distillation [8]. I would be keen to continue to work on bridging causality with neural networks.

Compositional and Grounded NLP Languages are compositional and grounded. However, benchmarks in the field mainly do not support rich grounding and learning compositional structures. Consequently, neural models trained with them often fail at simple adversarial tests involving compositionality.

I am excited about building compositional NLP systems, particularly in grounded environments. Humans can understand new linguistic phrases that are composed of learned ones. Nevertheless, most NLP systems are not compositional. There are several ways of understanding whether neural models store information in structured ways, including probing and causal analysis through interventions. Training neural models with constrained losses from probing and causal analysis may induce causal

structures to neural models and augment neural models with modularity and compositionality. On the other hand, groundings can provide rich learning signals for studying compositionality such as navigation, image-to-text, and vQA tasks.

To improve the compositionality of NLP systems in a grounded environment, I tried to build a pragmatic agent, using the Rational Speech Act (RSA) framework in our work presented at SCiL 2021 [9]. We induced Baysian priors using the speaker and listener models and created a reconstructor-based pragmatic speaker model that learns color generation via pragmatically grounding comparative modifiers. This approach increases zero-shot performance. I recently led a project of developing a benchmark for evaluating compositional generalization in a grounded environment, which has been accepted to the NeurIPS 2021 Datasets and Benchmarks Track [10]. In the paper, I showed that neural models often fail to generalize to new composites of known concepts. One direction I am excited about is building compositional and modular language models with strong generalizability. In particular, I am interested in applying such models in program induction and synthesis tasks.

Career Plans In truth, my background tells a slightly more complicated story. I was determined to pursue a Ph.D. in NLP after studying at Stanford and working in industry as a lead software engineer. My career objective is to become a professor to research and to teach in NLP. My positive experiences particularly inform this choice as a teaching assistant (TA) in college and as a team leader at work. As a TA, I led discussion sessions and office hours to engage students with coursework and supervised students with final capstone projects. As a lead engineer, I mentored junior team members to deliver industrial-leading products by turning research ideas into products. Pursuing a Ph.D. will enable me to grow my interests in research while also gaining further teaching and mentoring experience.

At Stanford, I am especially interested in the work of Professors Christopher Potts and Christopher Manning, and Percy Liang. I also appreciate that the Stanford NLP group is a dynamic cohort with scientists from computer science, linguistics and cognitive science – this tight integration is a unique strength. Following the work of these groups has led me to see a precise fit for my skills and interests at Stanford, and I am confident that it is an excellent place for me to pursue a Ph.D.

[1] Zhengxuan Wu, Xiyu Zhang, Tan Zhi-Xuan, Jamil Zaki, and Desmond C. Ong. Attending to emotional narratives. In *IEEE Affective Computing and Intelligent Interaction (ACII)*, 2019. https://arxiv.org/abs/1907.04197.

[2] Zhengxuan Wu, Thanh-Son Nguyen, and Desmond C. Ong. Structured self-attention weights encodes semantics in sentiment analysis. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 255–264, 2020. https://arxiv.org/abs/2010.04922.

[3] Zhengxuan Wu and Desmond C. Ong. On explaining your explanations of bert: An empirical study with sequence classification. *M.s., Stanford University*, 2021. https://arxiv.org/abs/2101.00196.

[4] Zhengxuan Wu and Desmond C. Ong. Context-guided bert for targeted aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. https://arxiv.org/abs/2010.07523.

[5] Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. DynaSent: A dynamic benchmark 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)*, pages 2388–2404, Online, August 2021. Association for Computational Linguistics. https://arxiv.org/abs/2012.15349.

[6] Zhengxuan Wu, Nelson F. Liu, and Christopher Potts. Identifying the limits of cross-domain knowledge transfer for pretrained models. 2021. https://arxiv.org/abs/2104.08410.

[7] Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing symbolic causal structures to produce systematic and interpretable neural networks. *M.s., Stanford University*, 2021. https://arxiv.org/abs/2112.00826.

[8] Zhengxuan Wu, Atticus Geiger, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Causal distillation for language models. *M.s., Stanford University*, 2021. https://zen-wu.social/papers/ACL22_CausalDistill.pdf.

[9] Zhengxuan Wu and Desmond C. Ong. Pragmatically informative color generation by grounding contextual modifiers. *Proceedings of the Society for Computation in Linguistics*, 4(1):438–445, 2021. https://arxiv.org/abs/2010.04372.

[10] Zhengxuan Wu, Elisa Kreiss, Desmond C. Ong, and Christopher Potts. ReaSCAN: Compositional reasoning in language grounding. *NeurIPS 2021 Datasets and Benchmarks Track*, 2021. https://arxiv.org/abs/2109.08994.