Language fascinates me as a tool for humans to refer to and reason about both the world around them and abstract concepts and theories. Drawn by this interest, I find recent language models an exciting object of study. I believe that a close analysis of the inner workings of these models can help with two important goals: (1) aligning their processing with existing theories of language to help models inherit useful human-like properties and (2) understanding the reasons behind the apparent success of these models, thereby expanding our knowledge of natural language. During my PhD, I hope to work on developing better tools for interpreting machine learning models and apply the resulting insight to improve their applications. This goal is roughly encapsulated by the following topics:

NLP and Philosophy of Language/Cognition My interest in language was initially sparked by two courses I took in my first year of college, and has since evolved into a broader and deeper examination of philosophy of language, metalogic, and consciousness. These fields attempt to theorize the complex nature of many capabilities any human being seems to have no issue mastering, such as the understanding of intention, having a theory of mind, and being able to refer to non-linguistic entities. On the other hand, it largely remains unknown if current language models are capable of having the same capabilities. A meaningful research direction would be to test and augment neural networks with existing theories of human linguistic capabilities.

I am fortunate to have received Brown's Undergraduate Research & Teaching Award, which supports me this Fall to conduct research in this direction. We build upon the idea of "meaning as use", where the meaning of an utterance is given by its use across situations under which it's produced. This is practically achieved by grounding sentence embedding of captions with their corresponding images. In addition to the standard next token prediction objective, our approach enforces these grounded sentences to be a "unit of meaning". We hypothesize that this helps models gain richer semantics content that conventional multimodal pretraining often fails to achieve [3]. I helped concretize the motivation into a specific training workflow, and am currently working on developing image encoders that capture relational structures, with the ultimate goal of transferring this information to a downstream language model.

I hope to continue working on drawing connections between theories of human capabilities and artificial models. As linguistic data is inherently sparse and only a proxy of the complex world we live in, I believe this line of work can help build models with more human-like properties, thus leading to better generalization and wider applicability.

Interpretability of Neural Networks While much work has been put into designing the "right" architecture and training schemes, it's often hard to control how neural networks learn and what solutions they learn. Hand-designing solutions, like in the project above that I worked on, is an iterative process that often fails or brings unintended consequences. I began to wonder if there are ways to reveal what solutions models learn and if such knowledge can help us control model behaviors. This sparked my interest in neural network interpretability.

As a first step, I worked at Brown's Language Understanding and Representation (LUNAR) Lab on a mechanistic interpretability project. Building upon previous work [2], we found that trained neural networks are often assembled modularly, with subnetworks capturing subtasks of their original training objective. We proposed a novel algorithm that localizes these subnetworks and transfers them to another model, thereby changing the solutions learned by the model to be more reliant on the transferred subnetwork. Our method shifts the inductive bias of models in

a controlled and sample-efficient manner, while also giving a finer grain of control. We applied it to address some well-known biases, such as the preference of local texture over global shape information of convolutional neural networks, and demonstrated that it works considerably better than data augmentation. This work is currently under review as a first-author paper at ICLR 2024 [4, 1].

I hope to continue to work on interpreting models and aligning their behaviors, possibly under the same framework. Much of the difficulty in this line of work lies in our lack of knowledge of a ground truth solution — in our work above, for example, we localize the subnetwork with an ad hoc dataset, which would be impossible to construct if the subtask cannot be defined. This motivates me to also work on other interpretability approaches that relax this constraint, such as through unsupervised training and direct analysis of representations.

Natural Language and Robotics During my work as an engineering intern creating a collaborative robot workstation automating production, I found that machine learning algorithms are brittle and lack modularity, and had to resort to hard-coding operations for reliability. Similar phenomena are quite common in engineering fields in general. The gap between the type of solutions we need and the ones we currently have became a motivating factor for me to pursue research. To narrow this gap, I see natural language as a useful tool to help induce complex world knowledge to various domain-specific applications. As an initial exploration, I worked on a project building a pipeline that used CLIP to provide semantic information to a robot's input feed and deployed it on Boston Dynamic's Spot robot. Our proof of concept is limited in capabilities, but it leads to a promising research direction that I'm excited to explore further: using natural language as an interface to interact with and provide information to physical agents.

Motivation to obtain a PhD I look forward to furthering my work on the above topics as well as exploring other relevant areas, with the long-term goal of leading a research career in either academia or industry. I see PhD studies as the best opportunity that enable this goal. Yale offers a vibrant research community that I'll be thrilled to be a part of. In particular, I hope to be advised by Professor **Tom McCoy**, whose research interests are closely aligned with mine. In addition, I also hope to work with Professor **Arman Cohan**, whose work spans both the science and applications of language models, and is also quite aligned to my interests.

[1] Anonymous. "Instilling Inductive Biases with Subnetworks". In: *Submitted to The Twelfth International Conference on Learning Representations*. under review. 2023. URL: https://openreview.net/forum?id=B4nhr6OJWI.

[2] Michael A. Lepori, Thomas Serre, and Ellie Pavlick. *Break It Down: Evidence for Structural Compositionality in Neural Networks*. 2023. arXiv: 2301.10884 [cs.CL].

[3] Tian Yun, Chen Sun, and Ellie Pavlick. *Does Vision-and-Language Pretraining Improve Lexical Grounding?* 2021. arXiv: 2109.10246 [cs.CL].

[4] Enyan Zhang, Michael A. Lepori, and Ellie Pavlick. *Instilling Inductive Biases with Subnetworks*. 2023. arXiv: 2310.10899 [cs.LG].