Pre-trained language models (LMs) with billions of parameters have proven their versatility as general-purpose NLP systems. By general-purpose, I mean that a single, frozen LM can be used in lieu of a task-specific architecture or fine-tuned model for a multitude of tasks. However, these models have unexpected failure modes and are notoriously difficult to control. Responsibly developing these models requires proper tools for understanding them. As an undergraduate computational linguistics researcher at Harvard (advised by Stuart Shieber and Yonatan Belinkov) and subsequently as a pre-doctoral researcher at AI2 (advised by Peter Clark and Ashish Sabharwal) I have cultivated a multi-perspective approach to understanding and using LMs. I aim to build on this work during my PhD by drawing on fields such as syntax and formal language theory to create practical and theoretical frameworks towards **understanding, evaluating, and safely using** language models as **general-purpose NLP systems.**

Understanding and improving LMs through linguistics Modern NLP is largely divorced from our understanding of linguistics and cognition in humans. Reconciling these two can lend clarity to our understanding of how LMs work, which in turn is the first step towards improving them. As my first foray into this effort, I set out to discover how modern LMs handle syntactic agreement. In my first-author ACL 2021 paper [1] we intervene on individual neurons in large transformers to observe their causal effect on syntactic agreement. We find, among other things, that transformers learn two distinct mechanisms for number agreement, and that these mechanisms are distributed in the activations of the network, rather than concentrated in any single model component, as was found with gender bias [2]. These findings constitute a step forward in understanding and interpreting how LMs mimic syntactic behavior, shedding light into the black box model. Subsequent research has built on our work in order to interpret other linguistic phenomena such as distributivity [3]. In future research, I hope to explore the potential of linguistically informed approaches to evaluate and improve the capabilities of LMs. As an example, I am interested in leveraging the language acquisition literature to evaluate and improve LMs' abilities to acquire new words by inferring their meaning from context. Improving this ability might be achieved via a self-supervised pre-training objective that randomly replaces lexical items with new unseen words, forcing the model to learn how to generalize effectively. Success here would result in LMs that are more robust to linguistic distribution shifts and adapt to evolving language.

Understanding LMs through formal languages I am excited by projects that borrow from formal language theory to increase our understanding of LMs. This approach makes it possible to answer questions that might otherwise be very difficult to study via natural language. For instance, I used regular languages to **measure the capabilities of transformers as instruction followers** in RegSet, my first-author EMNLP 2022 paper [4]. Large, pre-trained LMs can solve some NLP tasks by conditioning their generations on natural language instructions for the task [5, 6]. On the other hand, research has shown that neural models consistently struggle with compositional generalization [7]. This bodes poorly for the instruction following regime where the space of task descriptions is both intractably large and highly compositional. Moreover, the fuzziness of natural language makes it difficult to predict what types of instructions may be challenging for transformers. To solve this predicament, we propose a controllable proxy for studying instruction learning by studying LMs' ability to follow instructions in the form of regular expressions. We test the effects of attributes of regular languages, such as starfreeness [8], on their difficulty as instructions. Our experiments lead us to a number of intriguing hypotheses about what makes instruction learning hard, including evidence that even large transformers struggle with modular counting (e.g., determining whether something is even or odd). By taking advantage of the well studied attributes of formal languages, we achieve fine-grained control over our data, leading to findings that would have been extremely difficult and expensive to obtain on natural data. This approach can be applied more broadly to develop benchmarks that isolate and measure progress towards specific abilities in transformers that we might hope to see in natural language settings.

We can also use formal language theory to derive theoretical results that help us understand neural LMs. I am currently developing a framework for understanding what transformers can learn from instructions. In particular, I hope to show **which formal language families are provably learnable by transformers via instructions.** This builds on prior work [9], and provides both a principled way to study transformers, and bounds on what we can expect them to learn. In the future, I would like to study **the theoretical implications of sub-task decomposition.** I have previously worked on empirical studies in this area: in a preprint currently under submission to ICLR 2023 [10], I implement a modular method for recursively prompting large LMs which vastly improves generalization to longer sequence lengths for certain types of tasks when compared to other step-by-step reasoning styles. While exciting, it is not clear *why* these methods are effective. Recent work [11] proves that subtask decomposition via step-by-step reasoning enables learning difficult tasks via fine-tuning. Extending this work by characterizing the additional computational power intermediate reasoning affords transformers could explain why this prompting strategy has proven so successful.

Evaluating and mitigating risks from general-purpose LMs Comprehensive evaluation is critical to advancing general-purpose NLP systems, and existing benchmarks are often insufficient. For instance, existing datasets for are too narrow in scope to holistically evaluate math reasoning skills in LMs. To address this, I led a team of 11 researchers in compiling a diverse natural language **math reasoning benchmark** called LĪLA [12] (first-author paper, EMNLP 2022). We curate over 140K math problems and provide annotations for reasoning via program synthesis, Our experiments show that multitask learning and augmenting with a Python interpreter massively improves LM performance. Our multitask model, Bhāskara, outperforms similarly-sized models when fine-tuning on new math reasoning tasks. Despite our modeling contributions, LĪLA also shows that LMs are woefully deficient at math reasoning, and demonstrates the need for unified evaluations for aspiring general-purpose reasoning models. In my PhD I plan to continue to create thoughtful and comprehensive evaluations to measure progress towards general-purpose models.

During my PhD I also hope to expand research on the risks associated with general-purpose NLP systems and how to mitigate them. For instance, I am currently developing a decoding procedure for using frozen LMs as black boxes to generate their own task-specific prompts. I hope to apply this technique to study **adversarial prompts:** prompts that appear to elicit one behavior but cause the model to exhibit another. These prompts could be generated by simultaneously decoding for fluency for one task and accuracy on another. Exposing these types of vulnerabilities enables the research community to better achieve secure and ethical general-purpose NLP systems.

Fit for INSTITUTE I am specifically interested in joining the CS program at INSTITUTE because... Professor NAME's work on... is particularly interesting to me given my interest in...

References

[1] Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. Causal analysis of syntactic agreement mechanisms in neural language models. 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 1828–1843, Online, August 2021. Association for Computational Linguistics.

[2] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. Investigating gender bias in language models using causal mediation analysis. In *NeurIPS*, 2020.

[3] Pangbo Ban, Yifan Jiang, Tianran Liu, and Shane Steinert-Threlkeld. Testing pre-trained language models' understanding of distributivity via causal mediation analysis. *ArXiv*, abs/2209.04761, 2022.

[4] Matthew Finlayson, Kyle Richardson, Ashish Sabharwal, and Peter Clark. What makes instruction learning hard? an investigation and a new challenge in a synthetic environment. *EMNLP*, abs/2204.09148, 2022.

[5] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In *ACL*, 2022.

[6] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *ICLR*, 2022.

[7] Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *ICML*, 2018.

[8] Arto Salomaa. Jewels of formal language theory. page 53, 1981.

[9] William Merrill and Ashish Sabharwal. Log-precision transformers are constant-depth uniform threshold circuits. *ArXiv*, abs/2207.00729, 2022.

[10] Tushar Khot, H. Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. *ArXiv*, abs/2210.02406, 2022.

[11] Noam Wies, Yoav Levine, and Amnon Shashua. Sub-task decomposition enables learning in sequence to sequence tasks. *ArXiv*, abs/2204.02892, 2022.

[12] Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, and A. Kalyan. Lila: A unified benchmark for mathematical reasoning. *EMNLP*, abs/2210.17517, 2022.