Pretrained language models (LMs) with billions of parameters have proven their versatility as general-purpose NLP systems, meaning that they can often outperform task-specific models. However, they can have unexpected failure modes and are notoriously difficult to control. Responsibly developing general-purpose LMs requires creating 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 cultivated a multi-perspective approach to understanding and using LMs. I aim to build on this 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. Reconciling these can lend clarity to how LMs work, which is a first step towards improving them. As my first foray into research, I set out to discover how modern LMs handle syntactic agreement. In my first-author ACL 2021 paper [1] we intervene on neurons in transformers to observe their causal effect. 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 elucidate how LMs mimic syntactic behavior, and subsequent research has built on our work to interpret other linguistic phenomena such as distributivity [3]. In future research, I hope to explore linguistically informed approaches to evaluate and improve LM capabilities. E.g., 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 pretraining on augmented data that introduces hypothetical but plausible linguistic changes forcing the model to learn how to generalize effectively. Success would result in LMs that 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 enables answering questions that we cannot 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, pretrained LMs solve some NLP tasks by conditioning their generations on natural language instructions for the task [5, 6]. On the other hand, LMs often struggle with compositional generalization [7]. This bodes poorly for the instruction following regime where the space of instructions is highly compositional. Moreover, natural language fuzziness complicates predicting which instructions are challenging for transformers. In response, we propose a proxy for instruction learning by studying instructions in the form of regular expressions. We test the effects of regular language attributes such as star-freeness [8] on their difficulty as instructions. Our experiments yield multiple hypotheses for what makes instruction learning hard, including evidence that even large transformers struggle with modular counting (e.g., distinguishing even from odd). Here, well-studied attributes of formal languages afford us fine-grained control over the data, precipitating findings that would be difficult and expensive to obtain on natural data. I hope to apply this approach more broadly to isolate and measure progress towards other desirable abilities in LMs.

We can also use formal language theory to derive theoretical results that help us understand neural LMs. I am currently developing a framework for proving **which formal language families are learnable by transformers via instructions.** This builds on prior work [9], and will hopefully provide bounds on what we can expect transformers to learn from instructions. In the future, I would also like to study **the theoretical implications of sub-task decomposition.** I have previously worked on empirical studies in this area: in my ICLR 2023 submission [10] I implement a modular, recursive LM prompting method which vastly improves generalization to longer sequence lengths compared to other step-by-step reasoning styles.

In the future, I am interested in understanding why these methods work by formally characterizing the additional computational power that intermediate reasoning affords transformers.

Evaluating and mitigating risks from general-purpose LMs Comprehensive evaluation is critical to advancing general-purpose NLP systems. For instance, existing datasets are too narrow in scope to holistically evaluate math reasoning skills in LMs. To address this, I led 11 researchers in compiling a diverse natural language **math reasoning benchmark**, LĪLA [11] (first-author, EMNLP 2022). We curate over 140K math problems with annotations for reasoning via program synthesis. Our experiments show that multitask learning and augmenting with a Python interpreter massively improve LM performance. Despite our modeling contributions, LĪLA shows that LMs are woefully deficient at math reasoning, and demonstrates the need for unified evaluations of this sort. Going forward, I plan to continue creating thoughtful, comprehensive benchmarks for general-purpose models.

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

At Cornell I am especially interested in working with Professors Alexander Rush and Yoav Artzi, based on our overlap in research interests. In particular, I would be excited to work with Dr. Rush on building controllable NLP systems by developing our theoretical understanding of generative models and applying it to build principled controls to improve them. I also find Dr. Artzi's research approach compelling, and would be interested in working together to build NLP systems that adapt to novel contexts, especially by drawing on our understanding of pragmatics and linguistic change. I would also be excited to work with other faculty in Cornell's NLP group and collaborate with the NLP-adjacent groups such as CLab and C.Psyd.

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] 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.