Pre-trained language models (LMs) with billions of parameters are the go-to models for many NLP tasks. The sheer scale of their training corpora and parameter counts appears to endow them with the ability to compete with, and frequently outperform, task-specific models. LMs, however, can often have unexpected failure modes and are notoriously difficult to control. The responsible refinement and improvement of these models depends upon the development of the necessary tools and frameworks for understanding them. While an undergraduate at Harvard and subsequently as a pre-doctoral researcher at AI2, I have developed an approach of **drawing on fields such as syntax and formal language theory to develop practical and theoretical frameworks towards understanding and evaluating language models as general-purpose NLP systems.**

Understanding LMs through linguistics Modern NLP has become largely divorced from our understanding of linguistics and cognition in humans. Reconciling these two can lend clarity to our understanding of how LMs work. 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 and oral presentation [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 MLP activations of the network. This finding contrasts with earlier research that found that gender bias effects were concentrated in the attention heads of the models [2]. Subsequent research has built on our work in order to analyze other linguistic phenomena such as distributivity [3], and the effect of relative clauses on agreement [4]. Studies like these lend important insights into how these models work internally. During my PhD, I hope to explore the potential of this style of research to evaluate and improve the linguistic capabilities of neural networks. 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. This 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 approach using natural language. For instance, I used regular languages to **measure the capabilities of transformers as general instruction followers** in RegSet, my first-author EMNLP 2022 paper and oral presentation [5]. Large, pre-trained LMs can solve some NLP tasks by conditioning their generations on natural language instructions for the task [6, 7]. On the other hand, research has shown that neural models consistently to struggle with compositional generalization [8]. This bodes poorly for the instruction following regime where the space of task descriptions is both intractably large and highly compositional. Moreover, the complexity 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 starfreeness, 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.

On the theoretical side, I am currently developing a framework for understanding what transformers can learn from instructions. This builds on prior work [9] that achieves an initial lower bound for this problem. In my PhD, I am interested in deriving additional theoretical results that help us understand modern neural architectures through formal languages. For instance, recent work [10] proves that subtask decomposition, in the style of chain-of-thought reasoning [11], enables learning difficult sequence-to-sequence tasks. I would be interested in extending this work to characterize the additional computational power afforded to transformers via subtask decomposition. These types of results are not only intellectually interesting, but also provide both a principled way to study transformers and bounds on what we can expect them to learn.

Methods and risks for general-purpose LMs Some of my past and current work deals with methods for utilizing pre-trained language models as general-purpose solvers and the potential risks associated with this paradigm. In a preprint currently under submission to ICLR 2023 [12], I develop a modular prompting method for recursively prompting large LMs in order to vastly improve length generalization compared to few-shot and step-by-step reasoning style prompting. In my current work, I am engineering a decoding procedure for using frozen LMs as black boxes to generate their own task-specific prompts. I hope to apply this work 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. I hope to continue research into adversarial attacks such as this during my PhD in order to develop secure and ethical general-purpose NLP systems.

Evaluating general-purpose LMs A crucial part of developing general-purpose systems is evaluating them. As an example, current evaluation schemes fail to holistically evaluate general-purpose math reasoning skills in LMs because they are far too narrow in scope. Towards improving evaluation in this area, I led a team of 11 researchers in compiling a **comprehensive and diverse natural language math reasoning benchmark.** I introduce the benchmark, LĪLA [13], in my first-author EMNLP 2022 paper where we draw together a diverse set over 140K math problems. We additionally provide valuable annotations for mathematical reasoning via program synthesis, and show that multitask learning, combined with augmenting the model with a Python interpreter, massively improves LMs ability to do math reasoning with explicit reasoning steps. Our publicly released multitask model, christened Bhāskara, outperforms similarly-sized T5 and GPT-Neo models when fine-tuning on new mathematical reasoning tasks. LĪLA shows that LMs in their current form are woefully deficient when it comes to math reasoning, and highlights the need for these kinds of unified evaluations for aspiring general-purpose math reasoning models. During my PhD I hope to continue to develop thoughtful and comprehensive evaluations to measure and promote research into models with greater general utility.

Future plans After my PhD and postdoc I hope to become a PI at an institution where I can achieve autonomy in choosing my research directions while also pursuing my passion for mentorship teaching. Ideally this means a professorship at a research university. I value autonomy in research because I work best when I can focus my energy on projects that are intellectually interesting to me. I have also especially become aware of the importance of mentorship throughout my undergrad and time at AI2. I am passionate about promoting access to mentorship as a way to level the playing field for the next generation of researchers.

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, Stuart M. Shieber, Sebastian Gehrmann, Tal Linzen, and Yonatan Belinkov. Causal analysis of syntactic agreement mechanisms in neural language models. *ArXiv*, abs/2106.06087, 2021.

[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] Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. *ArXiv*, abs/2105.06965, 2021.

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

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

[7] 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.

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

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

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

[11] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903, 2022.

[12] 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.

[13] Matthew Finlayson, Swaroop Mishra, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, and Ashwin Kalyan. LĪLA: A unified benchmark for mathematical reasoning. 2022.