I first took an interest in computational linguistics while learning to speak Tagalog. Through the process, I learned a deep appreciation for the language's rich morphology, and the immense complexity of human language. My enthusiasm for studying language only increased as I began as a undergraduate researcher in NLP at Harvard under the mentorship of Stuart Shieber and Yonatan Belinkov at Harvard, and subsequently began as a pre-doctoral investigator at AI2 advised by Peter Clark and Ashish Sabharwal. These mentors, as well as my many amazing colleagues at both institutions, have inspired me to pursue a PhD in natural language processing (NLP).

Though some subfields of NLP research have advanced by leaps and bounds in the past few years, our understanding of these systems remains unsatisfactory. While transformers may get impressive results on, e.g., question answering, we do not know what the fundamental limits of what current architectures can compute or approximate, or what these models are doing internally. Do they learn syntactic rules? Can they generalize to new problems composed of known concepts? In my PhD I hope to answer some of these questions by **drawing on fields such as formal language theory and syntax to develop practical and theoretical frameworks** for understanding modern NLP methods, especially when it comes to **understanding compositional generalization** and using **pre-trained language models as general-purpose systems**.

**Understanding LMs through linguistics and formal language theory.** Modern NLP has become largely divorced from our understanding of linguistics and cognition in humans. I am interested in reconciling our understanding of both. 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 [2] on syntactic agreement. We find that transformers learn two distinct mechanisms for number agreement, depending on whether the relevant tokens are adjacent or not. 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]. During my PhD I hope to continue this line of research by exploring how LMs handle (or fail to handle) other linguistic phenomena. I am particularly interested in linguistic generalization.[1] For instance, perhaps causal analysis could help us understand how transformers handle and acquire new words at inference time by intervening on contextualized embeddings within the model. Strengthening our understanding of how LMs deal with well-studied linguistic phenomena can give insight into what inductive biases these models learn and how to make them better.[2] I also believe that insights from computational methods can contribute to the field of linguistics by empirically lower-bounding the computational ingredients for producing various phenomena that we see in human speech.[3]

As noted earlier, I am also excited by projects that borrow from computation and formal language theory to generate insights about LMs. This approach makes it possible to answer questions that might otherwise be very difficult to approach. 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]. However, the complexity of natural language makes difficult to predict what types of instructions may be beyond the grasp of Transformers. To solve this predicament, we propose a controllable proxy for studying instruction learning by studying LMs ability to learn interpret regular expressions and recognize their strings. 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). Subsequent work [8] has built

[1] You say you're especially interested in linguistic generalization but then you don't say much about that. You do have a section on generalization in the next page, but presumably that's something different? It's also confusing to See a section about that far from the first mention of generalization.

[2] A common motivation but hard to achieve in practice. If you have an example or a specific idea it could help.

[3] This is a confusing sentence. What do you mean?

on our formalization to obtain important theoretical results by using circuit complexity theory to formally characterize transformers ability to execute instructions in the form of a circuit. In my PhD, I hope to continue developing both theoretical and empirical frameworks for understanding neural networks through formal languages and computation theory. For instance, perhaps we can prove theoretical upper bounds for what transformers can learn from instructions, or characterize the additional computational power afforded to transformers when they are allowed to generate multiple intermediate tokens before producing an output.[4] On the empirical side, we can use formal languages to develop benchmarks to isolate and measure progress towards specific abilities in transformers that we hope to see in natural language settings. This approach takes advantage of the well studied properties of formal language and gives us fine-grained control over the data.

**Generalization** My work on RegSet bridges to another area of research I hope to continue in during my PhD: generalization. Research has shown that neural models consistently to struggle with compositional generalization [9]. This bodes poorly for models in succeeding on tasks like the instruction following regime I studied with RegSet where the space of task descriptions is both intractably large and highly compositional. I am interested in studying and developing models and methods that tackle these types of issues.

Some of my past and current work deals with generalization by attempting to develop methods for utilizing pre-trained language models to their full capacity as general problem solvers. In a preprint currently under submission to ICLR 2023 [10], 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. Currently, I am developing methods for using language models to decode their own task-specific prompts. One exciting application of this method would be to explore the possibility of adversarial prompts: prompts that appear to elicit one behavior but cause the model to exhibit another. Research into adversarial prompts is an important step towards to developing secure and ethical general-purpose NLP systems.

A crucial part of developing general-purpose systems is be able to evaluate 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 [11], 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 multi-task 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.

[4] See a paper along these lines by Yoav Levine and others, I think on a parity task.

**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] Judea Pearl. Direct and indirect effects. *Probabilistic and Causal Inference*, 2001.

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

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

[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] 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.