Motivation I first took an interest in linguistics while learning to speak Tagalog over the course of two years living in the Philippines. Up to that point I had never given much thought to language. I had scraped by in my high school Spanish class, but I only began to appreciate the complex relationship of language with culture, politics, and personal meaning once I had the chance to fully immerse myself in a language so distant from my native tongue. My enthusiasm for language has not diminished since that time and I am excited to pursue a PhD in natural language processing (NLP).

My educational and career goals are driven by three main factors. The first is my passion for understanding language and reasoning through developing computational models thereof. This passion stems from my own curiosity, as well as from my excitement for the large potential impact of NLP on society. NLP is already beginning to transform us, from enabling communication, to unlocking previously inaccessible insights from unstructured text. In particular, I am interested in studying language models as general-purpose reasoning systems and better understanding what they can learn.

Secondly, I am driven by my enthusiasm for mentorship and education. I have benefited from many amazing mentors, and I have had the honor of being a mentor and teacher to others, both in academic settings as a Harvard Teaching Fellow, and outside of academia as a leader in multiple organizations committed to making the outdoors more accessible. Good mentorship is key to the success and well-being of the next generation of scientists and thinkers. Furthermore, it can be an important avenue for those with underprivileged backgrounds to find their footing.

Lastly, I am driven by a sense of urgency for developing fair and safe AI tools. AI has great potential for good in our world, but we are already seeing that progress comes with significant risks, such as biased and toxic text generation. Fairness and ethics in AI are perhaps the most pressing issues we face as progress in NLP accelerates.

Development plans My immediate education goal is to obtain a PhD where I can develop my research vision; hone my mentorship, and teaching skills; and receive the guidance I need to fulfill my eventual goal of becoming a principal investigator (PI). I hope to become a PI at an institution where I can continue to do impactful research and continue to serve as a mentor to others.

Research experience My past research experiences have prepared me for graduate study by giving me significant experience in ownership, development, and execution of impactful research ideas. I have also gained important experience working with other researchers in both large and small-group research settings. In this section, I detail some of my research experiences thus far.

Though some subfields of NLP research have advanced by leaps and bounds in the past few years, our understanding of how these advances have come about 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? As I began my research career, I made the goal of answering these questions.

As my first step towards this goal, I set out to **interpret how modern language models handle syntactic agreement** while an undergrad at Harvard, advised by Stuart Shieber and Yonatan Belinkov. As a first author on this project, **I designed and ran compute-heavy experiments on a GPU cluster** to intervene on individual neurons in large transformers and observe the resulting effects. After running the experiments and obtaining results, we discovered that another group (Aaron Mueller advised by Sebastian Gehrmann and Tal Linzen) had independently ran very similar experiments and obtained almost identical results. In our co-authored ACL 2021 paper [**?**] we apply causal analysis to understand how transformers handle number agreement between nouns and verbs. We find that transformers generally learn two distinct mechanisms for agreement, depending on whether the relevant tokens are adjacent or not. Subsequent work has used causal

analysis for analyzing other linguistic phenomena such as distributivity [1], and the effect of relative clauses on agreement [**?**]. Strengthening our understanding of how (and whether) language models deal with well-studied linguistic phenomena can give insight into what inductive biases these models learn, and can guide research towards methods for improvement.

Excited by our findings, and starting as a pre-doctoral researcher at the Allen Institute for AI (AI2), I decided to next explore how another field, formal language theory, can increase our understanding of the capabilities of transformer models. In particular, I found formal language to be an excellent test-bed for studying how well language models **generalize on highly compositional tasks**. Compositional generalization is one area in which neural methods have been shown to consistently underperform humans [4]. This becomes particularly problematic as **instruction following** emerges as a prominent paradigm [6, 8] for building general-purpose large language models. Instruction following is where a language model is expected to perform a novel task (one not seen in training) given only a description of the task. Since this paradigm has emerged relatively recently little is known about what kinds of tasks current models can be expected to learn. To evaluate this and provide a synthetic sandbox for future research, I introduce RegSet [2] in my **first-author EMNLP 2022 paper** advised by Kyle Richardson, Ashish Sabharwal and Peter Clark. In our work we propose a highly controllable proxy for studying instruction learning by studying a language models ability to learn to interpret regular expressions. Regular expressions are succinct representations of regular languages, a well studied and highly compositional class of formal languages. As a result of our experiments, we develop a handful 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), less precise instructions, and tracking long contexts. We also release our challenging dataset, RegSet, to the public for further research. Subsequent work [5] has built on our ideas by using formal language theory and computation theory to characterize transformers ability to learn from instructions. I am currently working with Ashish Sabharwal to further expand this framework and **find stronger theoretical bounds on what transformers can learn from inputs**. My hope is that a better theoretical understanding of the limitations of transformers will inform our understanding of the architectural requirements for modeling language.

Building off of my work on learning from instructions, I am interested in **learning how to leverage and control language models beyond fine-tuning** by using techniques like in-context learning [**?**]. In our preprint [3], working closely with Tushar Khot and advised by Ashish Sabharwal, we develop methods for using large language models as problem decomposers and modular sub-problem solvers to improve performance over other prompting techniques such as chain-of-thought [7]. As a future direction, I am interested in developing a decoding method for using a language model to self-generate optimal task-specific prompts relying only on inference-time sequence probability estimates from the model. Additionally, I am interested in developing methods for inference-time adversarial attacks on large language models. This line of work has the potential to uncover both highly efficient techniques for leveraging large language models, and also uncover vulnerabilities that urgently need to be addressed in order to develop safe and fair AI systems.

While uncovering shortcomings of neural networks as general instruction followers with the RegSet work, I became interested in also evaluating neural networks as **general-purpose math reasoners**. In a contrasting approach, I led the effort to compile a **comprehensive natural language math reasoning benchmark** to evaluate language models' abilities on diverse math problem types. I introduce the benchmark, Līla [**?**], in my EMNLP 2022 paper (advised by Ashwin Kalyan and co-first-authored with Swaroop Mishra). Current evaluation schemes fail to comprehensively capture general-purpose math reasoning skills in language models because they are far too narrow in scope. As a result, they often overestimate the ability of particular models optimized for a single type of math reasoning. In our large scale effort we draw together a diverse set of mathematical tasks and unite them under a single benchmark of over 140K math problems. We provide valuable annotations for mathematical reasoning via program synthesis, where the language model has access to a Python interpreter. Our experiments show that **multi-task learning**, combined with

augmenting the model with a Python interpreter improves general-purpose math reasoning and the resulting model is an effective starting point for downstream fine-tuning. At the same time, our benchmark shows that language models, in their current form, are woefully deficient when it comes to math reasoning. **I led and contributed to all aspects of the Līla paper**, including collecting and annotating the datasets, the experimental design, running the experiments, and writing the paper. Accomplishing this meant **leading a team of 11 researchers**. My hope is that our high quality, comprehensive evaluation will serve to unify evaluation towards developing general-purpose math reasoning models.

Intellectual merit My research career thus far has yielded several significant results for the field. My initial work with causal analysis revealed interesting results for understanding how transformers handle syntactic agreement, forming a bridge between linguistics and modern NLP. My subsequent project on learning from instructions provided empirical evidence for important theoretical result on what transformers cannot learn, as well as provided a valuable tool and framework for studying instruction learning, an emerging paradigm in NLP. Līla provides a much needed comprehensive benchmark towards evaluating math reasoning models, a major improvement over previous, fragmented evaluations. Additionally, by releasing the models we developed with Līla, we provide future researchers with a strong starting point for further model development.

Broader impacts Beyond purely academic findings and technical technical improvements, my research has several broader societal implications.

Often, media and irresponsible research groups are incentivised to over-claim when it comes to the capabilities of new AI systems. This AI hype can be dangerous because it often leads to overzealous deployment of these systems and overconfidence in them. This in turn can lead to risks as these models may unexpectedly regurgitate unchecked biases and toxicity from their training data, or fail when it comes to mission-critical decision making. With both RegSet and Līla, our datasets point out major shortcomings in language models' reasoning and instruction-following abilities. These findings are important because they help combat AI hype, and ground claims about how well these systems actually work.

On the flip side, Līla represents an important step towards building reliable math reasoning systems. Such systems have significant potential for good, serving to assist and augment human achievements. For instance, a reliable AI assistant capable of complex mathematical reasoning could help a person struggling to understand and control their personal finances.

Conclusion I have spent the past few years pursuing my passion for NLP and computational linguistics by developing my research skills as an undergraduate at Harvard and pre-doctoral investigator at AI2. I am excited to pursue a PhD as the next chapter in my career and further develop my knowledge, skills, and leadership as a researcher.

References [1] Ban, P., et al. Testing pre-trained language models' understanding of distributivity via causal mediation [2] Finlayson, M., et al. What makes instruction learning hard? an investigation and a new challenge in a [3] Khot, T., et al. Decomposed prompting: A modular approach for solving complex tasks. ArXiv, [4] Lake, B. M. and Baroni, M. Generalization without systematicity: On the compositional skills of [5] Merrill, W. and Sabharwal, A. Log-precision transformers are constant-depth uniform threshold circuits. [6] Mishra, S., et al. Cross-task generalization via natural language crowdsourcing instructions. In ACL. [7] Wei, J., et al. Chain of thought prompting elicits reasoning in large language models. [8] Wei, J., et al. Finetuned language models are zero-shot learners. In ICLR. 2022.