I aspire to build language-driven agents capable of **robust reasoning**. Despite recent advances, models struggle to reason over scientific documents, reliably express their uncertainty, and update their world models. I have been fortunate to be advised by Prof. Percy Liang at Stanford and mentors at the Allen Institute for AI (AI2) along two directions for robust reasoning in large language models (LLMs) — **(1) consistency in LLM reasoning** and **(2) uncertainty quantification for LLMs**. During my PhD, I am eager to further explore these directions and more broadly study better LLM reasoning through areas such as long-horizon planning and continual learning.

Consistency in Reasoning. We found that LLMs were surprisingly inconsistent when *generating* a response and *validating* the same response. ChatGPT, for example, responds to ‘‘What is 7+8?’’ with ‘‘15’’, but says ‘‘No’’ when asked ‘‘Is 7+8=15?’’. Motivated to resolve this behavior, we bootstrapped a self-supervised signal of consistency from examples that were GV-consistent and fine-tuned models to penalize GV inconsistencies. This work, currently under review at ICLR 2024, produced models that were x% more consistent and more correct and gave me a taste for leveraging self-supervised approaches to correct model behavior. During my PhD, I would be excited to study the source of inconsistencies and hallucinations in LLMs. For instance, since training data can contain contradictory information, could better data curation lead to more consistent models? I am also interested in developing techniques to interpretably surface a model’s encoded knowledge and explicitly condition on it to reduce hallucinations.

Reliable Uncertainty Estimation. Standard methods to estimate uncertainty rely on output probabilities, which several state-of-the-art models (e.g. GPT-4, Claude) do not provide access to. To bridge this gap, I led a project to investigate if white-box models could act as surrogates to estimate the uncertainty of black-box models and found that Llama-2 can reliably approximate GPT-4 and Claude’s internal confidences. In a paper under review at ICLR 2024, I uncovered that confidence scores can transfer because *different* models tend to make *similar* mistakes. During my PhD, I would be curious to study the other properties that transfer between models and use them to further study black-box models through white-box models. Uncertainty estimation for longform generations, the challenging problem of extrapolating sequence level uncertainty from token level probabilities, is another area I would love to work on.

Career goals. My long term career goal is to lead a research group, in academia or industry, to work on solving the most pressing problems of the time using AI. **A PhD at UCLA** is the right next step toward my goals, as it will allow me to grow massively as a researcher, while exploring the problems that I deeply care about. At UCLA, I would be especially interested in working with Professors Kai-Wei Chang and Violet Peng. I see a strong fit for my research skills and aspirations at UCLA and I firmly believe that it is an ideal place for me to pursue my PhD.