I aspire to build language-driven agents capable of **robust reasoning**. Despite recent advances, current models struggle to reason over long-form scientific documents, ground their generations in factual knowledge, reliably express their uncertainty, and update their world models to incorporate new knowledge. I have been fortunate to be advised by Prof. Percy Liang at Stanford and amazing mentors at the Allen Institute for AI (AI2) along two directions that I believe are critical for making large language models (LLMs) robust reasoners:

1. **Consistency in Reasoning.** Models often contradict themselves in their generations, impeding their reliability. How can we encourage models to be consistent in their beliefs and generations?

2. **Uncertainty Quantification.** It is critical for users to know when they can trust model outputs. How can we build models that *know* what they don't know? How can we reliably quantify their uncertainty?

During my PhD, I would be excited to continue working on these directions and expand my research to other important aspects of reasoning, long-horizon planning, continual learning, and grounding responses in reliable sources to reduce hallucinations. My current work and career goals are detailed below.

Consistency in Reasoning. A robust and trustworthy AI agent should not only have the capability to tackle complex problems, but also maintain consistency in its beliefs and responses. At its core a lack of consistency can suggest a troubling lack of *understanding* — does a model that answers "Yes" to one surface form of a question and "No" to another really understand what is being asked?

LLMs can be remarkably inconsistent, especially while generating chain-of-thought (CoT) explanations — often contradicting themselves across generations, conversational turns, and sometimes within a single response. Motivated to resolve this behavior, I led a project to reduce inconsistencies in LLM reasoning chains. Neural models are black-box reasoners without any explicit mechanisms to detect and resolve inconsistencies in their generations. My intuitive approach was to add a more transparent reasoning layer on top of LLMs — using CoTs to surface model beliefs, prompting models to discover belief relationships, and using a constraint solver to identify a consistent set of beliefs. Although this approach achieved the goal of improving interpretability, it proved to be highly prompt-sensitive and challenging to generalize to the broad range of inconsistencies models struggle with. This propelled me to focus on examining specific types of inconsistencies and develop tailored solutions to resolve them.

A consistent model that generates a response should also believe in the correctness of its generation, a behavior we term "generator-validator (GV) consistency". At Stanford, we found that LLMs often fail to meet this basic consistency expectation. For example, we found that ChatGPT responds to "What is 7+8?" with "15", but says "No" when asked "Is 7+8=15?". We built upon our intuition that a self-supervised signal of consistency could be bootstrapped from examples where models were GV-consistent, and proposed *consistency fine-tuning*. In this simple and effective approach, we fine-tune models with an optimization objective penalizing generator-validator inconsistencies. In a paper currently under review [1], we show that our approach significantly reduced GV-inconsistency, demonstrating that novel training paradigms can be used to resolve inconsistencies in LLMs. Interestingly, we found that these improvements in consistency also translated to improvements in overall correctness — suggesting that resolving inconsistencies not only enables trust but can also be a gateway to better LLM performance.

During my PhD, I would be excited to systematically study the source of inconsistencies and hallucinations in LLMs, especially through the lens of training data and objectives. For instance, training data may contain contradictory information, causing models to learn a distribution over inconsistent beliefs. Could better data curation lead to more consistent models? Fine-tuning models on small task-specific datasets may be insufficient to fully adapt their parametric knowledge. Could this force models to make predictions beyond their internal knowledge and have the unintended consequence of teaching them to 'guess'? I would also be interested in developing interpretable techniques to inspect a model's encoded knowledge and ground its predictions in this knowledge to reduce hallucinations. One approach could be to surface and encode this latent knowledge into structured knowledge representations and explicitly condition on these structures.

Reliable Uncertainty Estimation. Trustworthy AI agents should know what they don't know and provide reliable uncertainty estimates alongside their predictions. This is especially important for domains like healthcare and law where confidently incorrect predictions could have significant negative implications.

Standard methods for confidence estimation rely on output probabilities from the model. However, increasingly state-of-the-art black-box models (e.g. GPT-4, Claude) do not provide access to model probabilities, making it difficult to trust their generations. Since open white-box models like Llama-2 provide output probabilities, I led a project to investigate if these white-box models could act as surrogates to estimate the uncertainty of black-box models. In a paper currently under review [2], we discovered that Llama-2 can reliably approximate GPT-4 and Claude's internal confidences, which is intriguing because it is a smaller, weaker model from a different model family. Driven to understand this surprising finding, we conducted careful analyses and discovered that *different* LLMs tend to make *similar* mistakes, in turn allowing confidence estimates to transfer between models. Our work capitalizes on this ingrained similarity in model behavior and provides a simple and reliable method to estimate black-box model confidences through white-box surrogate models. In future work, I am motivated to investigate what other properties transfer between models. How can we leverage these dimensions of transferability to study black-box models through the lens of white-box models? I am also excited to investigate which aspects of models most contribute to this observed transferability — the transformer architecture, pre-training objectives, internet-scale datasets, or fine-tuning methodology (instruction tuning, RLHF)?

Prompting models for their confidence in their answers is another natural means of eliciting confidence estimates from black-box models. We studied this behavior by linguistically probing LLMs ("What is your confidence on a scale of 1 to 5?") and found this not to work well [2]. Building upon the key intuition that it is likely easier to assess confidence in a contrastive setting ("Which of the following questions are you more confident about?"), in ongoing work, we found that contrastively probing the model yields more reliable confidence estimates. We are currently exploring ways to train small student models that can reliably mimic this contrastive signal from bigger models. We plan on using these small models as efficient uncertainty estimators for large models. These small models could also be used as reward models to align LLMs to produce more correct generations. We aim to submit this work to ICML 2024.

During my PhD, I would be excited to study uncertainty estimation for long-form generations. This is a challenging problem since it is difficult to extrapolate sequence-level uncertainty scores from token-level probabilities. Furthermore, models may associate different levels of uncertainty with different aspects of their generation, e.g. correctness vs completeness vs creativity. One interesting idea here could be to extend our contrastive confidence approach to ask models to contrastively assess uncertainty along different attributes of their generations and predict a confidence for each attribute.

Career goals. My long term career goal is to lead a research group — be it in academia or industry — and work on solving the most pressing problems of the time using AI. An exciting north star for my long term research, which ties in closely with my current work on robust reasoning with LLMs, is building agents which can serve as research assistants for novel scientific discoveries — learning from scientific documents, formulating grounded hypotheses, making experimentation plans, and updating plans based on empirical results. My future ambitions are driven by my past experiences building trustworthy and reliable language models at Stanford [1, 2], studying harmful biases in these models at AI2 [3], and developing efficient natural language processing (NLP) systems impacting millions of users at Microsoft [4, 5].

A PhD at Stanford 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. My experiences in mentoring students, as an intern mentor at Microsoft and a teaching assistant at Caltech, have been deeply rewarding and a PhD program would provide me the opportunity to explore this path further. The Stanford NLP group constantly redefines excellence in research through bold, groundbreaking work that seeks to fundamentally *question* assumptions and deeply *understand* NLP systems. These traits are deeply aligned with the research style I aspire to follow. At Stanford, I would be especially interested in working with Professors Percy Liang, Tatsu Hashimoto, Diyi Yang, and Chris Manning. Having worked with the amazing faculty and PhD students in the Stanford NLP group, I see a strong fit for my research skills and aspirations and I firmly believe that Stanford is an ideal place for me to pursue my PhD.

Publications

[1] Benchmarking and Improving Generator-Validator Consistency of Language Models.
X. Lisa Li, **V. Shrivastava**, S. Li, T. Hashimoto, P. Liang. 2023.
Under review
[ArXiV]

[2] Llamas Know What GPTs Don't Show: Surrogate Models for Confidence Estimation.
V. Shrivastava, P. Liang, A. Kumar. 2023.
Under review
[ArXiV]

[3] Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs.
S. Gupta, **V. Shrivastava**, A. Deshpande, A. Kalyan, P. Clark, A. Sabharwal, T. Khot. 2023.
Under review
[ArXiV]

[4] UserIdentifier: Implicit User Representations for Simple and Effective Personalized Sentiment Analysis.
F. Mireshghallah, **V. Shrivastava**, M. Shokouhi, T. Berg-Kirkpatrick, R. Sim, D. Dimitriadis. 2021.
In *North American Chapter of the Association for Computational Linguistics (NAACL) 2022*
[ArXiV]

[5] Exploring Low-Cost Transformer Model Compression for Large-Scale Commercial Reply Suggestions.
V. Shrivastava*, R. Gaonkar*, S. Gupta*, A. Jha. 2021.
Arxiv Pre-print
[ArXiV]