"Devise a ten-year plan to eradicate Malaria worldwide." My long-term vision is to build large language models (LLMs) capable of solving such complex problems. To achieve this ambitious goal, current LLMs need significant improvements in their **reasoning** abilities: 1) Consistency: LLMs should have consistent world models of the problem and prior studies 2) Uncertainty estimation: they should know what they don't know to communicate their uncertainty with stakeholders, 3) Human behavior simulation: these models should accurately predict public response to different proposals, 4) Long-horizon planning: they should create long-term plans that factor in the uncertainty of outcomes, resource constraints, etc. 5) Continual learning: lastly, LLMs should also be able to adapt their plans to shifting real-world signals.

I am excited to pursue a Ph.D. to work towards solving these problems. As a student researcher at the Allen Institute for AI (AI2), I worked on evaluating the fidelity of human behavior simulation through LLMs [3]. As a Master's student at Stanford advised by Prof. Percy Liang, my work has led to improvements in reasoning consistency [1] and uncertainty quantification [2] in LLMs, which I will outline below.

Improving Consistency in LLMs. Current LLMs demonstrate inconsistency in their beliefs and reasoning, e.g. we found that ChatGPT generates "15" when asked "What is 7+8?", but says "No" when asked to verify its own generation "Is 7+8=15?". We define this lack of consistency between the generation and validation modes of LLMs as 'generator-validator (GV) inconsistency'. This basic inconsistency in state-of-the-art LLMs highlights their lack of understanding and hampers their trustworthiness.

How can we resolve these inconsistencies? In [1], we proposed consistency fine-tuning, a novel fine-tuning method that builds upon 2 key observations: 1) current LLMs are trained on datasets containing inconsistencies and 2) current LLMs are not directly optimized for consistency. Our method addresses these limitations by bootstrapping a self-supervised dataset with no GV inconsistencies (i.e. where both generator and validator modes of the LLM agree) and fine-tuning the LLM on this dataset. This formulation encourages the generator to produce responses in agreement with the validator's signal and the validator to prefer the generator's outputs. **We show that our fine-tuned models not only demonstrate significant reductions in GV inconsistencies for our target tasks, but these improvements in consistency also extend to unseen tasks and domains.** Interestingly, we found this improved consistency to lead to an overall improvement in LLM accuracy, suggesting that consistency improvements could be used to unlock performance improvements.

During my Ph.D., I would be excited to further employ self-supervised approaches to correct other undesirable LLM behaviors, such as hallucinations. Our generator-validator approach is a specialized variant of the self-supervised paradigm called self-play where models improve by "playing against themselves". An interesting idea could be to develop collaborative self-play tasks where "lying" is implicitly penalized since it is disadvantageous for good collaboration. I am also excited to study the source of inconsistencies and hallucinations in LLMs through the lens of training data and objectives.

Uncertainty Estimation through Surrogate LLMs. Trustworthy AI agents should know what they don't know and provide reliable uncertainty estimates. Increasingly state-of-the-art black-box models (e.g. GPT-4, Claude) do not provide probability estimates, making it difficult to ascertain their uncertainty in their generations.

Can we approximate the internal confidences of black-box models using open white-box models? In [2], we show that **the uncertainty of answers from black-box models like GPT-4 and Claude can be reliably estimated through answer probabilities from open models like Llama 2**. Thus, our work provides a way to obtain high-quality answers from stronger black-box models, while using weaker white-box surrogate models to reliably estimate the uncertainty of those answers. To further understand this behavior, we conducted careful analyses and discovered that different LLMs tend to make similar mistakes, potentially enabling the transfer of their ingrained uncertainty.

Our work poses interesting questions that I wish to further explore: 1) What is the source of this uncertainty transfer: the shared architecture, pre-training objectives, datasets, or fine-tuning methods? and 2) What other properties transfer between models and how can we use them to study black-box models using white-box models? During my Ph.D., I would also like to study uncertainty estimation for long-form gen-

erations (such as summaries or plans), as simple extrapolation of token-level confidences to sequence-level confidences doesn't work well. One promising approach could be to assess the LLM's uncertainty along different aspects of the generation (e.g. correctness, completeness, or creativity).

Career goals. My future ambitions are driven by my past experiences studying consistency in LLM reasoning and better uncertainty estimation at Stanford [1, 2], evaluating the ability of LLMs to simulate human behavior at AI2 [3], and developing efficient natural language processing (NLP) systems impacting millions of users at Microsoft [4, 5]. My long-term career goal is to lead a research group solving the most pressing problems of the time using NLP. At MIT, I am keen to collaborate with **Professors Jacob Andreas, Yoon Kim, and Regina Barzilay**. I see a strong fit for my research skills and interests at MIT and I firmly believe that it is an ideal place for me to pursue my Ph.D.

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]
[Patent pending]

[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]