"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, prior studies, and any confounding factors involved, 2) *Uncertainty estimation*: they should know what they don't know to communicate their uncertainty with stakeholders and suggest appropriate actions (e.g. gathering new data), 3) *Human behavior simulation*: these models should accurately predict public response to different proposals to be able to formulate realistic plans, 4) *Long-horizon planning*: they should create long-term plans that factor in stakeholders, the uncertainty of outcomes, resource constraints, etc. 5) *Continual learning and adaptation*: 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 Master's student at Stanford advised by Prof. Percy Liang and as a student researcher at the Allen Institute for AI (AI2), my work has led to improvements in reasoning consistency [1] and uncertainty quantification [2] in LLMs, and has revealed that LLMs can perpetuate harmful biases while emulating the behaviors of various socio-demographic groups [3]. Below, I briefly outline my work in these three areas.

Improving Consistency in LLMs. Current LLMs demonstrate inconsistency in their beliefs and reasoning, e.g. we found that ChatGPT[1] 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 reliability and 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 outside the training domain.** Interestingly, we found this improved consistency to lead to an overall improvement in LLM accuracy, suggesting that consistency improvements are not only vital for reliability but could also 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 broad 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 training objectives.

Uncertainty Estimation through Surrogate LLMs. Trustworthy AI agents should know what they don't know and provide reliable uncertainty estimates for their predictions. This is especially important for domains like healthcare and policy where confidently incorrect predictions can have significant negative implications. Unfortunately, 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 confidences of such 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.

[1]Observed in the June 2023 version of gpt-3.5-turbo

Our work poses some interesting questions that I would be excited to explore in future research: 1) What is the primary source of this transferability: the shared architecture, pre-training objectives, datasets, or fine-tuning methodology? and 2) What other properties transfer between models and how can we utilize these similarities to study black-box models through the lens of white-box models? During my Ph.D., I would also like to develop better uncertainty estimators for long-form generations (such as summaries or plans), as simple extrapolation of token-level confidences to sequence-level confidences doesn't work as well. One promising approach could be to identify relevant task-specific aspects (e.g. correctness, completeness, or creativity) and assess the LLM's uncertainty along these different aspects of the generation.

Biased human behavior simulation by LLMs. The intriguing emergent ability of LLMs to adopt personas (*"Take on the role of a billionaire who supports universal basic income"*) offers the potential to simulate and study human behavior. However, the utility of such agents in making inferences about human behavior hinges on the fidelity with which they can simulate it.

How good are current LLMs at emulating people from different socio-demographic backgrounds? In my recent work in collaboration with AI2 [3], we show that **LLMs provide biased simulations of human behavior and exhibit stereotypical and harmful reasoning patterns while adopting personas of individuals from different socio-demographic groups**. For instance, ChatGPT frequently makes limiting and unfounded assumptions about a physically disabled person ("*As a physically disabled person, I can't move and thus I am not good at math.*"), leading to a 33% drop in performance on a benchmark of 24 reasoning datasets. We found this troubling behavior to be prevalent across datasets, socio-demographic groups, and LLMs. Our work serves as a cautionary tale that bias runs deep in current models and that using these models to simulate human behavior can have unintended and harmful consequences.

Interestingly, LLMs reject such stereotypes when probed directly (*"Are physically disabled people less skilled at math?"*), presenting another instance of inconsistency in LLMs, and suggesting that current alignment efforts fall short of mitigating these biases. During my Ph.D., I would be excited to develop techniques that can enable LLMs to faithfully emulate personas while not propagating such harmful undesirable stereotypes. One simple approach could be to additionally align models on persona-induced responses. I am also excited about using persona-assigned LLMs as a test bed for simulating real-world scenarios, such as predicting a community's response to new economic policies.

Career goals. My future ambitions are driven by my past experiences studying robust LLM reasoning at Stanford and AI2 [1, 2, 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 to solve the most pressing problems of the time using NLP. As a Ph.D. student, I would hope to continue building on my previous experiences as a teaching assistant during my undergraduate studies and as an intern mentor at Microsoft to mentor undergraduates in research and be a teaching assistant for several core classes.

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]