

Vaishnavi Shrivastava

CONTACT	Email: vaish.shrivastava@stanford.edu	Homepage: https://vshrivas.github.io/
KEYWORDS	LLM calibration, multi-hop reasoning, retrieval augmentation	
EDUCATION	Stanford University Master of Science, Computer Science Advisor: Percy Liang	2022 - 2024 (projected)
	California Institute of Technology (Caltech) Bachelor of Science, Computer Science	2015 - 2019 3.9/4.0
PUBLICATIONS	<p>[1] <u>Llamas Know What GPTs Don't Show: Surrogate Models for Confidence Estimation.</u> V. Shrivastava, P. Liang, A. Kumar. 2023. <i>In submission</i> [arxiv]</p> <p>[2] <u>Benchmarking and Improving Generator-Validator Consistency of Language Models.</u> X. Lisa Li, V. Shrivastava, S. Li, T. Hashimoto, P. Liang. 2023. <i>In submission</i> [arxiv]</p> <p>[3] <u>Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs.</u> S. Gupta, V. Shrivastava, A. Deshpande, A. Kalyan, P. Clark, A. Saharwal, T. Khot. 2023. <i>In submission</i> [arxiv]</p> <p>[4] <u>UserIdentifier: Implicit User Representations for Simple and Effective Personalized Sentiment Analysis.</u> F. Mireshghallah, V. Shrivastava, M. Shokouhi, T. Berg-Kirkpatrick, R. Sim, D. Dimitriadis. 2021. <i>North American Chapter of the Association for Computational Linguistics (NAACL) 2022</i> [arxiv]</p> <p>[5] <u>Exploring Low-Cost Transformer Model Compression for Large-Scale Commercial Reply Suggestions.</u> V. Shrivastava*, R. Gaonkar*, S. Gupta*, A. Jha. 2021. <i>Preprint</i> [arxiv]</p>	
RESEARCH EXPERIENCE	Research Assistant: <ul style="list-style-type: none">Stanford University: Advised by Percy Liang (Sep'22 - Current) <i>Themes: LLMs, Calibration, Reasoning</i>Allen Institute for AI: Advised by Tushar Khot, Peter Clark (Jun'23 - Current) <i>Themes: Reasoning, Persona-guided LLMs, Calibration</i>	
WORK EXPERIENCE	Applied Scientist: <ul style="list-style-type: none">Microsoft AI: Suggested Replies & Summarization (Sep'19 - Aug'22) <i>Themes: Dialog Systems, Model Compression, Personalization, Summarization</i> Software Engineering Intern: <ul style="list-style-type: none">Microsoft AI: Knowledge Mining and Graphs Group (Jul'18 - Sep'18) <i>Themes: Key-Phrase Extraction, Part-of-Speech Tagging, Email Search</i>Microsoft: Substrate Data Store Group (Jun'17 - Sep'17) <i>Themes: Multi-threading, Backend, Thread-Safe Caching</i>Dell-EMC: (Jun'16 - Sep'16) <i>Themes: Distributed Computing Algorithms, Concurrent Services</i>	
TEACHING EXPERIENCE	Teaching Assistant: <ul style="list-style-type: none">Caltech: Machine Learning & Data Mining, CS 155 (Jan'19 - Mar'19)Caltech: Database System Implementation, CS 122 (Jan'18 - Mar'18)	

SELECTED
RESEARCH
PROJECTS

Surrogate Models for Confidence Estimation

(Jul'23 - Sep'23)

Advisor: Percy Liang, Ananya Kumar - Stanford University

- SoTA models like GPT-4 and Claude don't provide access to their probabilities making it difficult to assess their confidences in their outputs. Prompting for confidences doesn't work well.
- We introduce surrogate model calibration - using a white-box surrogate like Llama-2 to approximate the internal confidences of a black-box model like GPT-4.
- Mixing surrogate probabilities and prompted confidences leads to further gains.

Implicit Reasoning Biases in Persona-Assigned LLMs

(Jun'23 - Sep'23)

Advisor: Tushar Khot, Ashish Sabarwal - Allen Institute for AI

- LLMs have deep-rooted biases which can be surfaced through personas.
- Performance on 24 reasoning tasks shows that models assigned personas of certain demographic groups may abstain or make more implicit reasoning errors, conforming with social stereotypes.

Improving Generator-Validator Consistency in LLMs

(Apr'23 - Jun'23)

Advisor: Percy Liang, Lisa Li - Stanford University

[\[arxiv\]](#)

- LLM are inconsistent with generative (What is 7+8?) vs validation queries (7+8=15, True/False?).
- We propose a fine-tuning scheme to improve generator-validator consistency and show accuracy improvements and performance transfer between generators and validators.

Belief Aggregation for Factually Correct Reasoning

(Sep'22 - Mar'23)

Advisor: Percy Liang - Stanford University

- We propose sampling chains-of-thought and extracting LLM's 'beliefs' from those chains.
- Beliefs can be composed and used to verify LLM's world model for factually correct reasoning.

Implicit Personalized User Representations

(Jul - Dec'21)

Microsoft Research

[\[arxiv\]](#)

- We investigate using uniformly distributed, non-trainable, user-specific prompts for user-personalization, instead of trainable embeddings, to circumvent periodically training embeddings per user.
- We demonstrate that we can outperform SOTA prefix-tuning based results on a suite of sentiment analysis by up to 13%, resulting in a paper.

Low-Cost Transformer Model Compression

(Jul - Nov'20)

Microsoft Search, Assistant and Intelligence

[\[arxiv\]](#)

- We experiment with low-cost methods to compress Transformer bi-encoder based reply suggestion system, reducing training and inference times by 42% and 35% respectively.
- Investigate how dataset size, pre-trained model use, and domain adaptation of the pre-trained model affected the performance of compression techniques.
- We discover that large-data settings allow low-cost techniques to be very effective in compressing pre-trained model based architectures. Insights led to a paper and a talk.

TALKS

"Supercharging Reply Suggestions: Model Compression Solutions and Insights from a Real-World Setting". Microsoft Machine Learning, AI and Data Science Conference (MLADS) 2021

SELECTED
LEADERSHIP
POSITIONS

- Corporate Vice President, *Caltech IEEE*
- Treasurer, *Caltech Society of Women Engineers*
- Secretary, *Caltech Robogals*

REFERENCES

Percy Liang, Associate Professor, Stanford University
Milad Shokouhi, Partner Applied Scientist, Microsoft
Dan Schwartz, Principal Applied Scientist, Microsoft
Donnie Pinkston, Lecturer, Caltech