Vaishnavi Shrivastava

Contact

Email: `vaish.shrivastava@stanford.edu` Homepage: `https://vshrivas.github.io/`

Keywords

Large language models, mathematical and commonsense reasoning, language model alignment

Education

Stanford University 2022 - 2024 (projected)
Master of Science, Computer Science
Advisor: Percy Liang

California Institute of Technology (Caltech) 2015 - 2019
Bachelor of Science, Computer Science **3.9/4.0**

Publications

[1] Benchmarking and Improving Generator-Validator Consistency of Language Models.
X. Lisa Li, **V. Shrivastava**, S. Li, T. Hashimoto, P. Liang. 2023.
International Conference on Learning Representations (ICLR) 2024 [arxiv]

[2] 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.
International Conference on Learning Representations (ICLR) 2024 [arxiv]

[3] Llamas Know What GPTs Don't Show: Surrogate Models for Confidence Estimation.
V. Shrivastava, P. Liang, A. Kumar. 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.
North American Chapter of the Association for Computational Linguistics (NAACL) 2022 [arxiv]
Patent pending [patent app]

[5] Exploring Low-Cost Transformer Model Compression for Large-Scale Commercial Reply Suggestions.
V. Shrivastava*, R. Gaonkar*, S. Gupta*, A. Jha. 2021.
**Equal Contribution*
Microsoft Journal of Applied Research (MSJar), 2021 [arxiv]

Research Experience

Research Assistant:

- **Stanford University**: Advised by Percy Liang *(Sep'22 - Current)*
  Themes: LLMs, Calibration, Reasoning
- **Allen Institute for AI**: Advised by Tushar Khot *(Jun'23 - Dec'23)*
  Themes: Reasoning, Persona-guided LLMs, Calibration

Work Experience

Applied Scientist:

- **Microsoft AI**: Suggested Replies & Summarization *(Sep'19 - Aug'22)*
  Themes: Dialog Systems, Model Compression, Personalization, Summarization

Software Engineering Intern:

- **Microsoft AI**: Knowledge Mining and Graphs Group *(Jul'18 - Sep'18)*
  Themes: Key-Phrase Extraction, Part-of-Speech Tagging, Email Search
- **Microsoft**: Substrate Data Store Group *(Jun'17 - Sep'17)*
  Themes: Multi-threading, Backend, Thread-Safe Caching
- **Dell-EMC**: *(Jun'16 - Sep'16)*
  Themes: Distributed Computing Algorithms, Concurrent Services

Technical Skills

Languages: *Proficient*: Python | *Intermediate*: Java | *Basic*: C++
Toolkits: PyTorch, Tensorflow, HuggingFace, Keras, AzureML

Teaching Experience

Teaching Assistant:

- **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 [arxiv]

- Models like GPT-4 and Claude do not provide access to their probabilities, making it difficult to assess their confidences. Linguistically asking them for confidences does not 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.
- Composing 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 [arxiv]

- Large language models (LLMs) have deep-rooted biases which can be surfaced through personas.
- Models assigned personas of marginalized demographic groups suffer from significant drops in reasoning performance on 24 challenging tasks, conforming to harmful stereotypical biases.

Improving Generator-Validator Consistency in LLMs *(Apr'23 - Jun'23)*

Advisor: Percy Liang, Lisa Li - Stanford University [arxiv]

- LLMs are inconsistent in their generator (What is 7+8?) and validator behaviors (Is 7+8=15?).
- We propose a fine-tuning objective to improve generator-validator consistency and show significant improvements in consistency and correctness that also generalize out of distribution.

Implicit Personalized User Representations *(Jul - Dec'21)*

Microsoft Research [patent app] [arxiv]

- We investigate using 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.
- We 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.

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*