YILUN ZHOU

CONTACT

Email: yilun@csail.mit.edu
Website: http://yilunzhou.github.io/

EDUCATION

Massachusetts Institute of Technology (MIT) *June 2019 - February 2023*
Doctor of Philosophy (Ph.D.) GPA: 5.0 / 5.0
Department of Electrical Engineering and Computer Science (EECS)
Advisor: Julie Shah
Thesis: Techniques for Interpretability and Transparency of Black-Box Models

Massachusetts Institute of Technology (MIT) *Sept 2017 - June 2019*
Master of Science (M.S.) GPA: 5.0 / 5.0
Department of Electrical Engineering and Computer Science (EECS)
Advisor: Julie Shah

Duke University *August 2013 - December 2016*
Bachelor of Science in Engineering (B.S.E.) GPA: 3.98 / 4.0
Department of Computer Science
Department of Electrical and Computer Engineering
Advisor: George Konidaris and Kris Hauser

RESEARCH

My research mainly aims to understand modern deep learning models, in particular large language models (LLMs). On the one hand, I assess the inner workings of these models by developing novel interpretability analyses. On the other hand, I curate novel benchmark tasks to assess diverse model behaviors, such their creativity, hallucination, reasoning, and safety. These two directions together lead to more fine-grained model insights, which allows for model improvement in a more dedicated and efficient manner.

WORK EXPERIENCE

Salesforce Research, Senior Research Scientist — February 2024 - Present

Amazon, Applied Scientist II — January 2023 - February 2024

Microsoft Research, Research Intern — May 2022 - August 2022

NVIDIA, Research Intern — May 2021 - September 2021

Facebook AI, Research Intern — May 2020 - September 2020

TALKS

- **Math AI Workshop @ ICML 2024**. How and Where Do LLMs Fail on Math? A Multi-Faceted Evaluation of LLM Math Reasoning Skills. July 2024.
- **Amazon**. Algorithmic and Societal Characteristics of Model Explanations. August 2023.
- **Brown University**. Correctness and Understandability of Model Interpretability Methods. June 2022.

- **University of Michigan** (Guest Lecture for EECS 692: Advanced AI). Methods and Evaluations for *Post Hoc* Model-Agnostic Local Explanations. March 2022.
- **Meta AI**. The Missing User Manual for Model Interpretability Methods: Evaluation, Comprehension and Integration. March 2022.
- **Future of Privacy Forum**. Model Explanations: Hopes, Setbacks and Paths Forward. February 2022.

SERVICE

Conference reviewer: ICML, NeurIPS, ICLR, AISTATS, AAAI, IJCAI, ACL, NAACL, EMNLP, CoRL, IROS

Journal reviewer: T-ASE, T-Cyb, IJHCI

Student volunteer: AISTATS 2021

Outstanding reviewer recognition: ICML 2022

PUBLICATIONS

Reverse chronological order. *Equal contribution.

J: journal. C: conference. W: workshop. P: pre-print. Highlighted work.

P3 **Yilun Zhou**, Caiming Xiong, Silvio Savarese and Chien-Sheng Wu. Shared Imagination: LLMs Hallucinate Alike. *arXiv preprint:2407.16604.*

P2 Daking Rai, **Yilun Zhou**, Shi Feng, Abulhair Saparov and Ziyu Yao. A Practical Review of Mechanistic Interpretability for Transformer-Based Language Models. *arXiv preprint:2407.02646.*

C13 Yujun Mao, Yoon Kim and **Yilun Zhou**. CHAMP: A Competition-level Dataset for Fine-Grained Analyses of LLMs' Mathematical Reasoning Capabilities. *Annual Meeting of the Association for Computational Linguistics (ACL) Findings, August 2024*

P1 Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, **Yilun Zhou** and Leilani H. Gilpin. Can Large Language Models Explain Themselves? A Study of LLM-Generated Self-Explanations. *arXiv preprint:2310.11207.*

W6 Shawn Im, Jacob Andreas and **Yilun Zhou**. Evaluating the Utility of Model Explanations for Model Development. *Under Review.*

C12 **Yilun Zhou**. Iterative Partial Fulfillment of Counterfactual Explanations: Benefits and Risks. *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*, August 2023.

C11 Daking Rai, Bailin Wang, **Yilun Zhou** and Ziyu Yao. Improving Generalization in Language Model-based Text-to-SQL Semantic Parsing: Two Simple Semantic Boundary-Based Techniques. *Annual Meeting of the Association for Computational Linguistics (ACL)*, July 2023.

C10 **Yilun Zhou** and Julie Shah. The Solvability of Interpretability Evaluation Metrics. *Conference of the European Chapter of the Association for Computational Linguistics (EACL) Findings*, May 2023.

W5 Daking Rai, **Yilun Zhou**, Bailin Wang and Ziyu Yao. Explaining Large Language Model-Based Neural Semantic Parsers. *AAAI Student Abstract and Poster Program*, February 2023.

C9 **Yilun Zhou**, Marco Tulio Ribeiro, and Julie Shah. ExSum: From Local Explanations to Model Understanding. *Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technology (NAACL-HLT)*, July 2022.

W4 Yiming Zheng, Serena Booth, Julie Shah, and **Yilun Zhou**. The Irrationality of Neural Rationale Models. *NAACL Workshop on Trustworthy Natural Language Processing (TrustNLP)*, July 2022.

C8 Ganesh Ghalme*, Vineet Nair*, Vishakha Patil*, and **Yilun Zhou***. Long-Term Resource Allocation Fairness in Average Markov Decision Process (AMDP) Environment. *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, May 2022.

J3 Mycal Tucker, **Yilun Zhou**, and Julie Shah. Latent Space Alignment Using Adversarially Guided Self-Play. *International Journal of Human-Computer Interaction (IJHCI)*, February 2022.

C7 **Yilun Zhou**, Serena Booth, Marco Tulio Ribeiro, and Julie Shah. Do Feature Attributions Correctly Attribute Features? *AAAI Conference on Artificial Intelligence (AAAI)*, February 2022.

C6 **Yilun Zhou**, Serena Booth, Nadia Figueroa, and Julie Shah. RoCUS: Robot Controller Understanding via Sampling. *Conference on Robot Learning (CoRL)*, November 2021.

C5 **Yilun Zhou**, Adithya Renduchintala, Xian Li, Sida Wang, Yashar Mehdad, and Asish Ghoshal. Towards Understanding the Behaviors of Optimal Deep Active Learning Algorithms. *Artificial Intelligence and Statistics (AISTATS)*, April 2021.

C4 Serena Booth*, **Yilun Zhou***, Ankit Shah, and Julie Shah. Bayes-TrEx: a Bayesian Sampling Approach to Model Transparency by Example. *AAAI Conference on Artificial Intelligence (AAAI)*, February 2021.

W3 Serena Booth*, Ankit Shah*, **Yilun Zhou***, and Julie Shah. Sampling Prediction-Matching Examples in Neural Networks: A Probabilistic Programming Approach. *AAAI Conference on Artificial Intelligence (AAAI) Workshop on Statistical Relational AI*, February 2020.

W2 **Yilun Zhou**, Julie Shah, and Steven Schockaert. Learning Household Task Knowledge from WikiHow Descriptions. *International Joint Conference on Artificial Intelligence (IJCAI) Workshop on Semantic Deep Learning*, August 2019.

C3 **Yilun Zhou**, Steven Schockaert, and Julie Shah. Predicting ConceptNet Path Quality Using Crowdsourced Assessments of Naturalness. *The Web Conference (WWW)*, May 2019.

J2 **Yilun Zhou**, Benjamin Burchfiel, and George Konidaris. Representing, Learning, and Controlling Complex Object Interactions. *Autonomous Robots (AuRo)*, April 2018.

W1 **Yilun Zhou** and Kris Hauser. 6DOF Grasp Planning by Optimizing a Deep Learning Scoring Function. *Robotics: Science and Systems (RSS) Workshop on Revisiting Contact - Turning a Problem into a Solution*, July 2017.

C2 **Yilun Zhou** and Kris Hauser. Incorporating Side-Channel Information into Convolutional Neural Networks for Robotic Tasks. *IEEE International Conference on Robotics and Automation (ICRA)*, May 2017.

J1 Kris Hauser and **Yilun Zhou**. Asymptotically Optimal Planning by Feasible Kinodynamic Planning in a State-Cost Space. *IEEE Transactions on Robotics (TRO)*, December 2016.

C1 **Yilun Zhou** and George Konidaris. Representing and Learning Complex Object Interactions. *Robotics: Science and Systems (RSS)*, June 2016.