

## Dylan Slack

Website: [dylanslacks.website](https://dylanslacks.website)  
Scholar: [scholar.google.com/dylanslack](https://scholar.google.com/dylanslack)

Email: [dslack@uci.edu](mailto:dslack@uci.edu)  
GitHub: [github.com/dylan-slack](https://github.com/dylan-slack)

<b>Bio</b>	I am a machine learning researcher with extensive experience in Large Language Models and RLHF. In my current role at Scale AI, I lead research efforts related to improving RLHF through human annotations, ensuring comprehensive evaluation of RLHF systems, and fundamental improvements to text representation learning. Previously, I complete a Ph.D. at UC Irvine, where I was co-advised across both UCI and Harvard University. My work on large language models has appeared at venues such as NeurIPS, Nature, ACL, and EMNLP. I generally prefer to work in PyTorch and Python.	
<b>Education</b>	<b>UC Irvine</b>	2019 - 2023
	<i>Ph.D. Computer Science</i> Advisors: Sameer Singh (UC Irvine), Himabindu Lakkaraju (Harvard University) <ul style="list-style-type: none"><li>• Dissertation: <i>Robust Interactions With Machine Learning Models</i></li></ul>	
	<b>Haverford College</b>	2015 - 2019
	<i>B.S. Computer Science (High Honors, Magna Cum Laude)</i> Advisor: Sorelle Friedler <ul style="list-style-type: none"><li>• Men's Varsity Lacrosse Team Captain</li></ul>	
<b>Employment</b>	<b>Scale AI</b>	2023 - Present
	<i>Machine Learning Research Engineer</i> Research Lead For: <ul style="list-style-type: none"><li>• <i>RLHF product</i></li><li>• <i>Text Representation Learning</i></li><li>• <i>Cheat Detection</i></li></ul>	
	<b>Google AI</b>	June 2021 - Sep. 2021
	<i>Research Intern</i> Research Advisors: Nevan Wichers, Yinlam Chow, Bo Dai <ul style="list-style-type: none"><li>• Developed generative models for constrained reinforcement learning, resulting in a publication.</li></ul>	
	<b>Amazon Web Services (AWS)</b>	June 2020 - Sep. 2020
	<i>Applied Scientist Intern</i> Research Advisors: Krishnaram Kenthapadi, Nathalie Rauschmayr	
	<b>DBO Partners</b>	Summer 2018
	<i>Summer Analyst</i>	
	<b>ValueAct Capital</b>	Summer 2017
	<i>Summer Analyst</i>	
	<b>SPMB</b>	Summer 2016
	<i>Summer Intern</i>	
<b>Selected Awards</b>	Honorable Mention Outstanding Paper, NeurIPS 2022 TSRML Workshop NeurIPS Outstanding Reviewer, 2021/2022 ICLR Outstanding Reviewer, 2021 Hasso Plattner Institute Fellow, 2021 (Full Ph.D. Funding)	
<b>Preprints</b>	TABLET: Learning From Instructions For Tabular Data. <b>Dylan Slack</b> , Sameer Singh. arXiv, 2023.	

- Technical Reports** A Holistic Approach For Test and Evaluation of Large Language Models. **Dylan Slack\***, Jean Wang\*, Denis Semenenko\*, Kate Park, Daniel Berrios, Sean Hendryx. 2023.
- Publications** Post Hoc Explanations of Language Models Can Improve Language Models. Satyapriya Krishna, Jiaqi Ma, **Dylan Slack**, Asma Ghandeharioun, Sameer Singh, Himabindu Lakkaraju. NeurIPS, 2023.
- TalkToModel: Understanding Machine Learning Models With Open Ended Dialogues. **Dylan Slack**, Satyapriya Krishna, Hima Lakkaraju\*, and Sameer Singh\*. Nature Machine Intelligence, 2023.
- Reliable Post hoc Explanations: Modeling Uncertainty in Explainability. **Dylan Slack**, Sophie Hilgard, Sameer Singh, and Hima Lakkaraju. NeurIPS, 2021.
- Counterfactual Explanations Can Be Manipulated. **Dylan Slack**, Sophie Hilgard, Hima Lakkaraju, and Sameer Singh. NeurIPS, 2021.
- Active Meta-Learning for Predicting and Selecting Perovskite Crystallization Experiments. Venkateswaran Shekar, Gareth Nicholas, Mansoor Ani Najeeb, Margaret Zeile, Vincent Yu, Xiaorong Wang, **Dylan Slack**, Zhi Li, Philip Nega, Emory Chan, Alexander Norquist, Joshua Schrier, Sorelle Friedler. The Journal of Chemical Physics, 2021.
- On the Lack of Robust Interpretability of Neural Text Classifiers. Muhammad Bilal Zafar, Michele Donini, **Dylan Slack**, Cedric Archambeau, Sanjiv Das, Krishnaram Kenthapadi. Findings of ACL, 2021.
- Context, Language Modeling, and Multimodal Data in Finance. Sanjiv Das, Connor Goggins, John He, George Karypis, Sandeep Krishnamurthy, Mitali Mahajan, Nagpur-nanand Prabhala, **Dylan Slack**, Rob van Dusen, Shenghua Yue, Sheng Zha, Shuai Zheng. The Journal of Financial Data Science, 2021.
- Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. **Dylan Slack\***, Sophie Hilgard\*, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES), 2020.
- Fairness Warnings and Fair-MAML: Learning Fairly with Minimal Data. **Dylan Slack**, Sorelle Friedler, and Emile Givental. ACM Conference on Fairness, Accountability and Transparency (FAccT), 2020.
- Workshop** Rethinking Explainability as a Dialogue: A Practitioner’s Perspective. Himabindu Lakkaraju\*, **Dylan Slack\***, Yuxin Chen, Chenhao Tan, and Sameer Singh, NeurIPS HCAI Workshop, 2022.
- SAFER: Data-Efficient and Safe Reinforcement Learning via Skill Acquisition. **Dylan Slack**, Yinlam Chow, Bo Dai, and Nevan Wichers, ICML DARL Workshop, 2022.
- Defuse: Training More Robust Models through Creation and Correction of Novel Model Errors. **Dylan Slack**, Nathalie Rauschmayr, Krishnaram Kenthapadi. NeurIPS XAI 4 Debugging Workshop 2021.
- Feature Attributions and Counterfactual Explanations Can Be Manipulated. **Dylan Slack**, Sophie Hilgard, Sameer Singh, and Himabindu Lakkaraju. ICML Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI, 2021.
- Differentially Private Language Models Benefit from Public Pre-training. Gavin Kerrigan\*, **Dylan Slack\***, and Jens Tuyls\*. EMNLP PrivNLP Workshop, 2020.

Assessing the Local Interpretability of Machine Learning Models. **Dylan Slack**, Sorelle A. Friedler, Carlos Scheidegger, and Chitradeep Dutta Roy. NeurIPS Workshop on Human-Centric Machine Learning, 2019.

## **Presentations**

### **Invited Talks & Presentations**

- Stanford MedAI, 2022. *Exposing Shortcomings and Improving the Reliability of Machine Learning Explanations*
- Imperial College, 2022. *Exposing Shortcomings and Improving the Reliability of Machine Learning Explanations*
- Meta, 2022. *Exposing Shortcomings and Improving the Reliability of Machine Learning Explanations*
- UCI CML, 2022. *Exposing Shortcomings and Improving the Reliability of Machine Learning Explanations*
- Harvard University, 2021. *Reliable Post Hoc Explanations*
- Aggregate Intellect, 2021. *Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods*

## **Patents**

Automatic Failure Diagnosis and Correction in Machine Learning Models  
Nathalie Rauschmayr, Krishnaram Kenthapadi, and **Dylan Slack**  
*Patent Application Filed*

## **Academic Service**

### **Community**

- KDD Deep Learning Day, Organizer, 2021.

### **Program Committee Member**

- NeurIPS (2019, 2020, 2021, 2022), FAccT (2021), ICLR (2021), ICML (2020), AAAI (2020, 2021), KDD (2019).