

Eric Wallace

E-mail: ericwallace@berkeley.edu
GitHub: github.com/Eric-Wallace
Web: ericswallace.com

EDUCATION	UC Berkeley Ph.D. in Computer Science GPA: 4.0/4.0	2019 - Present
	University of Maryland B.S. in Computer Engineering GPA: 3.9/4.0, GRE: 170/170Q, 168/170V, 6/6W	2014 - 2018
RESEARCH EXPERIENCE	UC Berkeley (Berkeley NLP, RISE, BAIR) <i>Research Assistant</i> Advisors: Dan Klein, Dawn Song	Berkeley, California Aug 2019 - Present
	Allen Institute for Artificial Intelligence (AI2) <i>Research Intern</i> Advisors: Matt Gardner, Sameer Singh	Irvine, California Jan 2019 - July 2019
	University of Maryland, CLIP Lab <i>Undergraduate Research Assistant</i> Advisor: Jordan Boyd-Graber	College Park, MD Jan 2018 - Dec 2018
INDUSTRY EXPERIENCE	Lyft, Self Driving Team <i>Software Engineering Intern</i>	Palo Alto, California June - Aug 2018
	Intel <i>Software Engineering Intern</i>	Folsom, California Aug - Dec 2017
	Appian <i>Software Engineering Intern</i>	Reston, Virginia May - Aug 2017
FELLOWSHIPS, AWARDS & HONORS	AI2 Intern of the Year, 2019 EMNLP Best Demo Award, 2019 EMNLP Travel Award 2018 EMNLP Best Reviewer Award, 2018 AIAA Student Conference Best Paper, 2017 Lockheed Martin Corporate Partners Scholarship, 2017 Yurie/Jeong H. Kim Scholarship, 2016 Leidos Corporate Partners Scholarship, 2016 University of Maryland Presidential Scholarship, 2014 Eagle Scout, 2012	

- [1] Pretrained Transformers Improve Out-of-Distribution Robustness
Dan Hendrycks*, Xiaoyuan Liu*, **Eric Wallace**, Adam Dziedziec, Rishabh Krishnan,
and Dawn Song.
Association for Computational Linguistics (ACL), 2020.
- [2] Universal Adversarial Triggers for Attacking and Analyzing NLP
Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh.
Empirical Methods in Natural Language Processing (EMNLP), 2019.
- [3] AllenNLP Interpret: A Framework for Explaining Predictions of NLP Models
Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh.
Demo at Empirical Methods in Natural Language Processing (EMNLP), 2019.
Best Demo Award
- [4] Do NLP Models Know Numbers? Probing Numeracy in Embeddings
Eric Wallace*, Yizhong Wang*, Sujian Li, Sameer Singh, and Matt Gardner.
Empirical Methods in Natural Language Processing (EMNLP), 2019.
- [5] Misleading Failures of Partial-input Baselines
Shi Feng, **Eric Wallace**, and Jordan Boyd-Graber.
Association for Computational Linguistics (ACL), 2019.
- [6] Compositional Questions Do Not Necessitate Multi-hop Reasoning
Sewon Min*, **Eric Wallace***, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi,
and Luke Zettlemoyer.
Association for Computational Linguistics (ACL), 2019.
- [7] Understanding Impacts of High-Order Loss Approximations and Features in
Deep Learning Interpretation
Sahil Singla, **Eric Wallace**, Shi Feng, and Soheil Feizi.
International Conference in Machine Learning (ICML), 2019.
- [8] Trick Me If You Can: Human-in-the-loop Generation of Adversarial Examples
for Question Answering
Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber.
Transactions of the Association for Computational Linguistics (TACL), 2019.
- [9] Pathologies of Neural Models Make Interpretations Difficult
Shi Feng, **Eric Wallace**, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez,
and Jordan Boyd-Graber.
Empirical Methods in Natural Language Processing (EMNLP), 2018.
- [10] Interpreting Neural Networks With Nearest Neighbors
Eric Wallace*, Shi Feng*, and Jordan Boyd-Graber.
EMNLP Workshop on Analyzing and Interpreting Neural Networks (BlackboxNLP), 2018.

TEACHING EXPERIENCE	EMNLP 2020 Tutorial - <i>Interpreting Predictions of NLP Models</i> Sameer Singh, Matt Gardner, Eric Wallace A tutorial on interpretability methods for NLP, e.g., saliency maps, input perturbations (LIME, input reduction, Anchors), and adversarial attacks (SEARs, universal adversarial triggers).	November 2020
MENTORING	Tony Zhao (2020-Present), UC Berkeley Undergraduate. Albert Xu (2020-Present), UC Berkeley Undergraduate. Nikhil Kandpal (2019-Present), Independent Researcher. Published [2]. Now PhD Student at UNC. Jens Tuyls (2019-2020), UC Irvine Undergraduate. Published [3]. Now PhD Student at Princeton. Junlin Wang (2019-2020), UC Irvine Undergraduate. Published [3]. Now Research Assistant at UC Irvine.	
TALKS	November 2019. <i>Universal Adversarial Triggers for Attacking and Analyzing NLP</i> . Empirical Methods in Natural Language Processing (EMNLP) in Hong Kong. November 2018. <i>Pathologies of Neural Models Make Interpretation Difficult</i> . Empirical Methods in Natural Language Processing (EMNLP) in Brussels, Belgium. March 2018. <i>Generalization in Deep Learning for Language</i> . Adobe Labs & UMD Computer Science Advisory Board in College Park, MD. November 2017. <i>Learning Macro-Based RL Policies</i> . DeepMind/Blizzard StarCraft AI Workshop in Anaheim, CA.	
ACADEMIC SERVICE	Program Committee Member <ul style="list-style-type: none"> • Association for Computational Linguistics (ACL): 2020 • Empirical Methods in Natural Language Processing (EMNLP): 2020, 2019, 2018 (<i>Best Reviewer Award</i>). • Workshop on NLP for Positive Impact: 2020 • International Workshop on Semantic Evaluation (SemEval): 2018 	
OPEN SOURCE SOFTWARE	AllenNLP (Contributor) A software library with abstractions for NLP research, written on top of PyTorch. Developer of the AllenNLP Interpretation Toolkit [3] (<i>EMNLP 2019 Best Demo</i>).	
PRESS & MEDIA	Universal Adversarial Triggers for Attacking and Analyzing NLP [2], Twitter , Wired , qbitai , Synced , NLP Newsletter . AllenNLP Interpret: A Framework for Explaining Predictions of NLP Models [3], Twitter , InfoQ , UC Irvine , NLP Newsletter Do NLP Models Know Numbers? Probing Numeracy in Embeddings [4] Twitter . Trick Me If You Can: Human-in-the-loop Generation of Adversarial Examples for Question Answering [8], Front page of Reddit , Dukakis Shaping Futures , UMD Press Release , UMD Podcast , AI2 NLP Highlights Podcast . Pathologies of Neural Models Make Interpretations Difficult [9]. AI2 NLP Highlights Podcast , TWiML Talk Podcast , UCI NLP , UMD . Interpreting Neural Networks with Nearest Neighbors [10]. UCI NLP	