

E-mail: ericwallace@berkeley.edu
Scholar: scholar.google.com/ericwallace
Twitter: twitter.com/Eric_Wallace_
Website: ericwallace.com

Eric Wallace

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 <i>Research Assistant</i> Advisors: Dan Klein, Dawn Song	Berkeley, California Aug 2019 - Present
	Facebook AI Research (FAIR) <i>Research Intern</i> Advisors: Robin Jia, Douwe Kiela	Menlo Park, California June 2021 - Sept 2021
	Allen Institute for Artificial Intelligence (AI2) <i>Research Intern</i> Advisors: Matt Gardner, Sameer Singh	Irvine, California Jan 2019 - Aug 2019
	University of Maryland <i>Undergraduate Research Assistant</i> Advisor: Jordan Boyd-Graber	College Park, MD Jan 2018 - Dec 2018
SWE 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
REFEREED PUBLICATIONS	<ul style="list-style-type: none">[1] Calibrate Before Use: Improving Few-shot Performance of Language Models Tony Z. Zhao*, Eric Wallace*, Shi Feng, Dan Klein, Sameer Singh <i>International Conference in Machine Learning (ICML)</i>, 2021.[2] Extracting Training Data from Large Language Models Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, Colin Raffel <i>USENIX Security Symposium</i>, 2021.[3] Concealed Data Poisoning Attacks on NLP Models Eric Wallace*, Tony Z. Zhao*, Shi Feng, Sameer Singh <i>North American Chapter of the Association for Computational Linguistics (NAACL)</i>, 2021.[4] Detoxifying Language Models Risks Marginalizing Minority Voices Albert Xu, Eshaan Pathak, Eric Wallace, Maarten Sap, Suchin Gururangan, Dan Klein <i>North American Chapter of the Association for Computational Linguistics (NAACL)</i>, 2021.[5] Imitation Attacks and Defenses for Black-box Machine Translation Systems Eric Wallace, Mitchell Stern, Dawn Song <i>Empirical Methods in Natural Language Processing (EMNLP)</i>, 2020.	

- [6] Evaluating Models’ Local Decision Boundaries via Contrast Sets
Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, **Eric Wallace**, Ally Zhang, Ben Zhou
Findings of Empirical Methods in Natural Language Processing (EMNLP Findings), 2020.
- [7] AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts
Taylor Shin*, Yasaman Razeghi*, Robert L Logan IV*, **Eric Wallace**, Sameer Singh
Empirical Methods in Natural Language Processing (EMNLP), 2020.
- [8] Gradient-based Analysis for NLP Models is Manipulatable
Junlin Wang*, Jens Tuyls*, **Eric Wallace**, Sameer Singh
Findings of Empirical Methods in Natural Language Processing (EMNLP Findings), 2020.
- [9] Train Large, Then Compress: Rethinking Model Size for Efficient Training and Inference of Transformers
Zhuohan Li*, **Eric Wallace***, Sheng Shen*, Kevin Lin*, Kurt Keutzer, Dan Klein, Joseph E. Gonzalez
International Conference in Machine Learning (ICML), 2020.
- [10] Pretrained Transformers Improve Out-of-Distribution Robustness
Dan Hendrycks*, Xiaoyuan Liu*, **Eric Wallace**, Adam Dziedzic, Rishabh Krishnan, Dawn Song
Association for Computational Linguistics (ACL), 2020.
- [11] Universal Adversarial Triggers for Attacking and Analyzing NLP
Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, Sameer Singh
Empirical Methods in Natural Language Processing (EMNLP), 2019.
- [12] AllenNLP Interpret: A Framework for Explaining Predictions of NLP Models
Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, Sameer Singh
Demo at Empirical Methods in Natural Language Processing (EMNLP), 2019.
Best Demo Award
- [13] Do NLP Models Know Numbers? Probing Numeracy in Embeddings
Eric Wallace*, Yizhong Wang*, Sujian Li, Sameer Singh, Matt Gardner
Empirical Methods in Natural Language Processing (EMNLP), 2019.
- [14] Misleading Failures of Partial-input Baselines
Shi Feng, **Eric Wallace**, Jordan Boyd-Graber
Association for Computational Linguistics (ACL), 2019.
- [15] Compositional Questions Do Not Necessitate Multi-hop Reasoning
Sewon Min*, **Eric Wallace***, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, Luke Zettlemoyer
Association for Computational Linguistics (ACL), 2019.
- [16] Understanding Impacts of High-Order Loss Approximations and Features in Deep Learning Interpretation
Sahil Singla, **Eric Wallace**, Shi Feng, Soheil Feizi.
International Conference in Machine Learning (ICML), 2019.
- [17] Trick Me If You Can: Human-in-the-loop Generation of Adversarial Examples for Question Answering
Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, Jordan Boyd-Graber
Transactions of the Association for Computational Linguistics (TACL), 2019.
- [18] Pathologies of Neural Models Make Interpretations Difficult
Shi Feng, **Eric Wallace**, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, Jordan Boyd-Graber
Empirical Methods in Natural Language Processing (EMNLP), 2018.

TEACHING EXPERIENCE	<p>EMNLP 2020 Tutorial—<i>Interpreting Predictions of NLP Models</i></p> <p>Eric Wallace, Sameer Singh, Matt Gardner</p> <p>A tutorial on interpretability methods for NLP, e.g., saliency maps, input perturbations, influence functions, and adversarial attacks.</p>	November 2020
MENTORING	<p><u>Tony Z. Zhao</u> (2020-2021), UC Berkeley Undergrad. Published [1, 3]. Now PhD student at Stanford.</p> <p><u>Albert Xu</u> (2020-2021), UC Berkeley Undergrad. Published [4]. Now PhD student at USC.</p> <p><u>Jens Tuyls</u> (2019-2020), UC Irvine Undergrad. Published [8,12]. Now PhD student at Princeton.</p> <p><u>Junlin Wang</u> (2019-2020), UC Irvine Undergrad. Published [8,12]. Now Masters student at UCI.</p> <p><u>Nikhil Kandpal</u> (2019), UMD Undergrad. Published [11]. Now PhD at UNC.</p>	
TALKS	<p><i>What Can We Learn from Vulnerabilities of NLP Models?</i> Stanford, Cornell, DeepMind, UT Austin, CMU.</p> <p><i>Conference Oral Presentations:</i> ICML 2021 Virtual [1], NAACL 2021 Virtual [3], EMNLP 2020 Virtual [5], ICML 2020 Virtual [9], ACL 2020 Virtual [10], EMNLP 2019 Hong Kong [11], EMNLP 2018 Brussels [18].</p>	
ACADEMIC SERVICE	<p>Program Committee Member</p> <ul style="list-style-type: none"> • ACL Rolling Review: 2021, 2022 • North American Chapter of the Association for Computational Linguistics (NAACL): 2021, 2022 • Association for Computational Linguistics (ACL): 2020, 2021, 2022 • International Conference on Machine Learning (ICML): 2021 • Neural Information Processing Systems (NeurIPS): 2020, 2021 • Empirical Methods in Natural Language Processing (EMNLP): 2018, 2019, 2020, 2021 • Workshops: RobustML Workshop (ICLR 2021), MRQA (EMNLP 2021), NLP for Positive Impact (ACL 2021), SRW (NAACL 2021), DistShift (NeurIPS 2021) <p>Student Volunteer</p> <ul style="list-style-type: none"> • Empirical Methods in Natural Language Processing (EMNLP): 2018 	
SELECTED PRESS & MEDIA	<p>Automated Crossword Solving, <u>Wired</u>, <u>Slate</u>, <u>BBC</u>, Science Friday, Top-10 of Hacker News, The Register, Le Big Data (French), <u>Berkeley Engineering Magazine</u>, <u>Daily Californian</u>, <u>WNPR</u>, <u>Sydney Morning Herald</u>, <u>NVIDIA</u>, <u>Neil deGrasse Tyson Podcast</u></p> <p>Extracting Training Data from Large Language Models [2], <u>Twitter #1</u>, <u>Twitter #2</u>, <u>Twitter #3</u>, <u>Google Blog</u>, <u>BAIR Blog</u>, <u>Nature News</u>, <u>Henry AI Labs</u>, <u>Wired</u>, <u>Yannic Kilcher</u>, <u>Top of Hacker News</u>, <u>Top of ML Reddit</u>, <u>Sebastian Ruder Highlights</u>.</p> <p>Train Large, Then Compress: Rethinking Model Size for Efficient Training and Inference of Transformers [9], <u>Twitter</u>, <u>TWiML Talk Podcast</u>, <u>Sebastian Ruder Highlights</u>, <u>Towards Data Science</u>, <u>Henry AI Labs Video</u>, <u>BAIR Blog</u>, <u>Sebastian Ruder Newsletter</u></p>	