Eric Wallace

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

Education

UC Berkeley 2019 - 2024 (expected)
Ph.D. in Computer Science
Research Advisors: Dan Klein, Dawn Song

University of Maryland 2014 - 2018
B.S. in Computer Engineering
GPA: 3.9, GRE: 170/170Q, 168/170V, 6/6W
Research Advisor: Jordan Boyd-Graber

Industry Experience

Facebook AI Research (FAIR) Menlo Park, California
Research Intern June 2021 - Sept 2021
Research Advisors: Robin Jia, Douwe Kiela

Allen Institute for Artificial Intelligence (AI2) Irvine, California
Research Intern Jan 2019 - Aug 2019
Research Advisors: Matt Gardner, Sameer Singh

Lyft, Self Driving Team Palo Alto, California
Software Engineering Intern June 2018 - Aug 2018

Intel Folsom, California
Software Engineering Intern Aug 2017 - Dec 2017

Awards & Honors

Apple Fellowship in AI/ML
Best Poster Award, NeurIPS 2021 ENLSP Workshop
Best Demo Award, EMNLP 2019
AI2 Intern of the Year, 2019
Eagle Scout, 2012

Publications

[1] Deduplicating Training Data Mitigates Privacy Risks in Language Models
Nikhil Kandpal, **Eric Wallace**, Collin Raffel
International Conference on Machine Learning (ICML), 2022.

[2] Automated Crossword Solving
Eric Wallace*, Nicholas Tomlin*, Albert Xu*, Kevin Yang*, Eshaan Pathak*, Matt Ginsberg, Dan Klein
Association for Computational Linguistics (ACL), 2022.

[3] Analyzing Dynamic Adversarial Training Data in the Limit
Eric Wallace, Adina Williams, Robin Jia, Douwe Kiela
Findings of the Association for Computational Linguistics (ACL Findings), 2022.

[4] Cutting Down on Prompts and Parameters: Simple Few-Shot Learning with Language Models
Robert L. Logan IV, Ivana Balažević, **Eric Wallace**, Fabio Petroni, Sameer Singh, Sebastian Riedel
ACL Findings 2022; NeurIPS Efficient NLP Workshop.
Best Poster Award

[5] Calibrate Before Use: Improving Few-shot Performance of Language Models
Tony Z. Zhao*, **Eric Wallace***, Shi Feng, Dan Klein, Sameer Singh
International Conference on Machine Learning (ICML), 2021.

[6] 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
USENIX Security Symposium, 2021.

[7] Concealed Data Poisoning Attacks on NLP Models
Eric Wallace*, Tony Z. Zhao*, Shi Feng, Sameer Singh
North American Chapter of the Association for Computational Linguistics (NAACL), 2021.

[8] Detoxifying Language Models Risks Marginalizing Minority Voices
Albert Xu, Eshaan Pathak, **Eric Wallace**, Maarten Sap, Suchin Gururangan, Dan Klein
North American Chapter of the Association for Computational Linguistics (NAACL), 2021.

[9] Imitation Attacks and Defenses for Black-box Machine Translation Systems
Eric Wallace, Mitchell Stern, Dawn Song
Empirical Methods in Natural Language Processing (EMNLP), 2020.

[10] 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 the Empirical Methods in Natural Language Processing (EMNLP Findings), 2020.

[11] 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.

[12] Gradient-based Analysis for NLP Models is Manipulable
Junlin Wang*, Jens Tuyls*, **Eric Wallace**, Sameer Singh
Findings of the Empirical Methods in Natural Language Processing (EMNLP Findings), 2020.

[13] 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 on Machine Learning (ICML), 2020.

[14] 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.

[15] 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.

[16] 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

[17] 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.

[18] Misleading Failures of Partial-input Baselines
Shi Feng, **Eric Wallace**, Jordan Boyd-Graber
Association for Computational Linguistics (ACL), 2019.

[19] 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.

[20] Understanding Impacts of High-Order Loss Approximations and Features in Deep Learning Interpretation
Sahil Singla, **Eric Wallace**, Shi Feng, Soheil Feizi.
International Conference on Machine Learning (ICML), 2019.

[21] 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.

[22] 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

Tutorials:
- EMNLP, 2020. *Interpreting Predictions of NLP Models.*

Guest Lectures:
- University of Minnesota CSCI 8980-06. *Robustness in NLP*
- UC Berkeley CS 288. *Robustness in NLP*
- University of Stuttgart. *Interpreting Predictions of NLP Models*

Panels:
- Women in Machine Learning. *PhD Fellowships Applications*
- ACL Mentoring. *How to Keep Up with Work in the Field*

Mentoring

Research Mentoring
- Tony Zhao (2020-2021), UC Berkeley Undergrad. Published [5, 7]. Now PhD student at Stanford.
- Albert Xu (2020-2021), UC Berkeley Undergrad. Published [2, 8]. Now PhD student at USC.
- Eshaan Pathak (2020-2021), UC Berkeley Undergrad. Published [2, 8]. Now at You.com
- Jens Tuyls (2019-2020), UC Irvine Undergrad. Published [12,16]. Now PhD student at Princeton.
- Junlin Wang (2019-2020), UC Irvine Undergrad. Published [12,16]. Now PhD student at Duke.
- Nikhil Kandpal (2019), UMD Undergrad. Published [15]. Now PhD student at UNC.

External Mentoring
- Women in Machine Learning. PhD Application Mentor
- Berkeley Equal Access Assistance Program. PhD Application Mentor
- Berkeley AI4All 2022. Instructor
- DEFCON 2022. AI Security Village Competition Judge

Presentations

Invited Talks & Presentations
- USC ISI, 2022. *Emerging Vulnerabilities in Large-scale NLP Models*
- Malicious Life Podcast. *Hacking Language Models*
- Stanford, 2021. *What Can We Learn from Vulnerabilities of NLP Models?*
- Cornell, 2021. *What Can We Learn from Vulnerabilities of NLP Models?*
- DeepMind, 2021. *What Can We Learn from Vulnerabilities of NLP Models?*
- UT Austin, 2021. *What Can We Learn from Vulnerabilities of NLP Models?*
- CMU, 2021. *What Can We Learn from Vulnerabilities of NLP Models?*

Conference Oral Presentations: ACL 2022 Dublin [2], ICML 2021 Virtual [5], NAACL 2021 Virtual [7], EMNLP 2020 Virtual [9], ICML 2020 Virtual [13]; ACL 2020 Virtual [14], EMNLP 2019 Hong Kong, [15], EMNLP 2018 Brussels [22].

Academic Service

Program Committee Member
- Conferences: ACL (2020, 2021, 2022), ICML (2021), NeurIPS (2020, 2021), EMNLP (2018, 2019, 2020, 2021, 2022), ACL Rolling Review (2021, 2022), ICLR (2023), NAACL (2021, 2022)
- Journals: TMLR (2022)
- Workshops: Distribution Shifts (NeurIPS 2022), Principles of Distribution Shifts (ICML 2022), BlackBox NLP (EMNLP 2022), RobustML Workshop (ICLR 2021), MRQA (EMNLP 2021), NLP for Positive Impact (ACL 2021), SRW (NAACL 2021), DistShift (NeurIPS 2021)

Selected Media & Press

Automated Crossword Solving [2], Discover, Wired, Slate, BBC, Science Friday, Top of Hacker News, The Register, Le Big Data (French), Berkeley Engineering Magazine, Daily Californian, WNPR, Sydney Morning Herald, NVIDIA, Neil deGrasse Tyson Podcast

Extracting Training Data from Large Language Models [6], Twitter #1, Twitter #2, Twitter #3, Google Blog, BAIR Blog, Nature News, Henry AI Labs, Wired, Yannic Kilcher, Top of Hacker News, Top of ML Reddit, Sebastian Ruder Highlights.

Train Large, Then Compress: Rethinking Model Size for Efficient Training and Inference of Transformers [13], Twitter, TWiML Talk Podcast, Sebastian Ruder Highlights, Towards Data Science, Henry AI Labs Video, BAIR Blog, Sebastian Ruder Newsletter