Eric Wallace

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

Education

UC Berkeley 2019 - 2023
Ph.D. in Computer Science
Research Advisors: Dan Klein, Dawn Song
Thesis: *Measuring and Mitigating Vulnerabilities of Language Models*

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

Google Brain Mountain View, CA
Research Intern June 2023 – Sep 2023
Research Advisors: Dustin Tran, Denny Zhou, Xinyun Chen

Facebook AI Research (FAIR) Menlo Park, CA
Research Intern June 2021 – Sep 2021
Research Advisors: Robin Jia, Douwe Kiela

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

Selected Awards

PET Award for Outstanding Research in Privacy, Runner Up 2023
Apple Fellowship in AI/ML, 2022-2024
Best Poster, NeurIPS 2021 ENLSP Workshop
Best Paper, EMNLP 2019 Demo Track
AI2 Intern of the Year, 2019
Eagle Scout, 2012

Publications

[1] The False Promise of Imitating Proprietary LLMs
Arnav Gudibande*, **Eric Wallace***, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, Dawn Song
arXiv preprint, 2023.

[2] Extracting Training Data from Diffusion Models
Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramér, Borja Balle, Daphne Ippolito, **Eric Wallace**
USENIX Security Symposium, 2023.

[3] Poisoning Instruction-Tuned Language Models
Alexander Wan*, **Eric Wallace***, Sheng Shen, Dan Klein
International Conference on Machine Learning (ICML), 2023.

[4] Large Language Models Struggle to Learn Long-Tail Knowledge
Nikhil Kandpal, Haikang Deng, Adam Roberts, **Eric Wallace**, Colin Raffel
International Conference on Machine Learning (ICML), 2023.

[5] InCoder: A Generative Model for Code Infilling and Synthesis
Daniel Fried, Armen Aghajanyan, Jessy Lin, Sida Wang, **Eric Wallace**, Freda Shi, Ruiqi Zhong, Wen-tau Yih, Luke Zettlemoyer, Mike Lewis
International Conference on Learning Representations (ICLR), 2023.
Spotlight Presentation (Top 25%)

[6] Measuring Forgetting of Memorized Training Examples
Matthew Jagielski, Om Thakkar, Florian Tramèr, Daphne Ippolito, Katherine Lee, Nicholas Carlini, **Eric Wallace**, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, Chiyuan Zhang
International Conference on Learning Representations (ICLR), 2023.

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

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

[9] 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.

[10] 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

[11] 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.
Long Oral Presentation (Top 3%)

[12] 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.
Runner up for PET Award (Outstanding Work in Privacy)

[13] 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.

[14] 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.

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

[16] Evaluating Models' Local Decision Boundaries via Contrast Sets
Matt Gardner, Yoav Artzi, ... (other authors hidden) ... **Eric Wallace**, Ally Zhang, Ben Zhou
Findings of the Empirical Methods in Natural Language Processing (EMNLP Findings), 2020.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

[28] 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

Courses:

- Co-instructor of Berkeley's graduate-level NLP (CS 288) with 90 students in Spring 2023. Taught alongside Dan Klein and Kevin Lin. I developed and taught ∼10 new lectures on language models and advanced NLP topics (e.g., RLHF, retrieval, vision-language models). I also developed new homeworks, coding assignments, and mentored students.
- TA for Berkeley's CS188: Artificial Intelligence. Summer 2023.

Tutorials:

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

Guest Lectures for Courses:

- Stanford CS 329X, 2023. *Security & Privacy in NLP*
- Washington University in St. Louis CSE 527A, 2022. *Security & Privacy in NLP*
- University of Minnesota CSCI 8980-06, 2022. *Robustness in NLP*
- UC Berkeley CS 288, 2022. *Robustness in NLP*
- ML @ Berkeley, 2022. *Security & Privacy in NLP*
- University of Stuttgart, 2022. *Interpreting Predictions of NLP Models*

Panels:

- Women in Machine Learning, 2022. *PhD Fellowships Applications*
- ACL Mentoring, 2022. *How to Keep Up with Work in the Field*
- Berkeley AI Hackathon (1500 participants), 2023. *Future of LLMs—Beyond Hacking*
- USENIX PEPR, 2023. *Privacy Challenges and Opportunities in LLM-Based Chatbots*

Mentoring

Student Research Mentoring

- Carolyn Wang (2023-Present), UC Berkeley Undergrad.
- Alex Wan (2022-Present), UC Berkeley Undergrad. Published [3].
- Arnav Gudibande (2022-2023), UC Berkeley Masters. Published [1]. Now at PerplexityAI.
- Tony Zhao (2020-2021), UC Berkeley Undergrad. Published [11, 13]. Now PhD at Stanford.
- Albert Xu (2020-2021), UC Berkeley Undergrad. Published [8, 14]. Now PhD student at USC.
- Eshaan Pathak (2020-2021), UC Berkeley Undergrad. Published [8, 14]. Now at You.com
- Jens Tuyls (2019-2020), UC Irvine Undergrad. Published [18,22]. Now PhD student at Princeton.
- Junlin Wang (2019-2020), UC Irvine Undergrad. Published [18,22]. Now PhD student at Duke.
- Nikhil Kandpal (2019), UMD Undergrad. Published [21]. Now PhD student at UNC.

Masters Thesis Review

- Arnav Gudibande. Chair: Dawn Song. 2023

Other Mentoring

- Women in Machine Learning (WiML), 2022–2023. PhD Application Assistance.
- Berkeley Equal Access Assistance Program (EAAA), 2022-2023. PhD Application Assistance.
- Berkeley AI4All, 2022. Instructor
- BAIR Undergraduate Mentoring, 2022–2024.

Presentations

Invited Talks

- Oracle Labs, 2023. *Memorization in Large Language Models*
- Princeton, 2023. *Memorization in Large Language Models*
- UMD, 2023. *Memorization in Large Language Models*
- UNC, 2023. *Memorization in Large Language Models*
- USC ISI, 2022. *Emerging Vulnerabilities in Large-scale NLP Models*

- Malicious Life Podcast, 2022. *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 [8], ICML 2021 Virtual [11], NAACL 2021 Virtual [13], EMNLP 2020 Virtual [15], ICML 2020 Virtual [19]; ACL 2020 Virtual [20], EMNLP 2019 Hong Kong, [21], EMNLP 2018 Brussels [28].

Academic Service

Program Committee Member

- Conferences: ACL (2020, 2021, 2022), ICML (2021, 2023), NeurIPS (2020, 2021, 2023), EMNLP (2018, 2019, 2020, 2021, 2022), ACL Rolling Review (2021, 2022), ICLR (2023), NAACL (2021, 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),

Departmental Service

- Berkeley PhD Admissions. 2021–2023
- Berkeley Student Committee for Faculty Hiring. 2023
- Berkeley PhD Visit Days Recruitment. 2021–2023

Selected Media & Press

Extracting Training Data from Diffusion Models [2], Twitter #1 (3 million views), Twitter #2, Twitter #3, MIT Technology Review, TWIML Podcast, Gizmodo, Ars Technica, Vice, TechSpot, The Register, New Scientist

Automated Crossword Solving [8], Discover, New Scientist, Wired, Slate, BBC, Science Friday, Top of Hacker News, The Register, Berkeley Engineering Magazine, WNPR, Daily Californian, NVIDIA Blog, Neil deGrasse Tyson Podcast

Extracting Training Data from Large Language Models [12], Twitter #1, Twitter #2, Twitter #3, Google Blog, BAIR Blog, Nature News, Henry AI Labs, MIT Technology Review, Wired, Yannic Kilcher Video, Top of Hacker News