Eric Wallace

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

Education

UC Berkeley 2019–2024
Ph.D. in Computer Science
Research Advisors: Dan Klein, Dawn Song
Thesis: *Emerging Vulnerabilities of Large Language Models*

University of Maryland 2014–2018
B.S. in Computer Engineering
Research Advisor: Jordan Boyd-Graber

Work Experience

OpenAI San Fransisco, CA
Member of Technical Staff Nov 2023–Present

Google Deepmind Mountain View, CA
Research Intern June 2023–Aug 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

Apple Fellowship in AI/ML, 2022–2024
Outstanding Paper Award, NeurIPS 2023 RegML Workshop
Runner up for PET Award (a test of time award for [19]), 2023
Best Poster, NeurIPS 2021 ENLSP Workshop
First Superhuman Crossword AI, ACPT 2021
Best Demo Paper, EMNLP 2019
AI2 Intern of the Year, 2019

Publications

[1] The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions
Eric Wallace*, Kai Xiao*, Reimar Leike*, Lilian Weng, Johannes Heidecke, Alex Beutel
arXiv preprint, 2024.

[2] Stealing Part of a Production Language Model
Nicholas Carlini, Krishnamurthy Dvijotham, Milad Nasr, A. Feder Cooper, Katherine Lee, Matthew Jagielski, Thomas Steinke, Daniel Paleka, Jonathan Hayase, Arthur Conmy, David Rolnick, Florian Tramér, **Eric Wallace**
International Conference on Machine Learning (ICML), 2024.

[3] Covert Malicious Finetuning: Subverting LLM Safety Training Without Detection
Danny Halawi*, Alexander Wei*, **Eric Wallace**, Tony Tong Wang, Nika Haghtalab, Jacob Steinhardt
International Conference on Machine Learning (ICML), 2024.

[4] What Evidence Do Language Models Find Convincing?
Alex Wan, **Eric Wallace**, Dan Klein
Association for Computational Linguistics (ACL), 2024.

[5] Scalable Extraction of Training Data from (Production) Language Models
Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, **Eric Wallace**, Florian Tramér, Katherine Lee
arXiv preprint, 2023.

[6] Privacy Side Channels in Machine Learning Systems
Edoardo Debenedetti, Giorgio Severi, Nicholas Carlini, Christopher A. Choquette-Choo, Matthew Jagielski, Milad Nasr, **Eric Wallace**, Florian Tramér
arXiv preprint, 2023.

[7] SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore
Sewon Min*, Suchin Gururangan*, **Eric Wallace**, Hannaneh Hajishirzi, Noah A. Smith, Luke Zettlemoyer

International Conference on Learning Representations (ICLR), 2024.
Spotlight Presentation (Top 5%)

[8] The False Promise of Imitating Proprietary LLMs
Arnav Gudibande*, **Eric Wallace***, Charlie Snell*, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, Dawn Song
International Conference on Learning Representations (ICLR), 2024.
Spotlight Presentation (Top 5%)

[9] 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.

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

[11] 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.

[12] 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

[13] 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.

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

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

[16] 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.

[17] 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

[18] 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%)

[19] 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 (Test of Time Award)

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.

[35] 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.
- Teaching assistant for Berkeley's CS188: Artificial Intelligence in Summer 2023. Aside from typical TA duties (e.g., leading discussion sections), I co-designed the midterm and final exam.

Tutorials:

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

Guest Lectures for Courses:

- UC Berkeley 194/294-267, 2024. *Memorization in Large Language Models*
- USC CSCI 699, 2023. *Security & Privacy in NLP*
- UC Berkeley CS294/194-196, 2023. *Intro & Foundations of LLMs*
- 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*

Mentoring

Student Research Mentoring

- Alex Wan (2022-2024), UC Berkeley Undergrad. Published [10].
- Carolyn Wang (2023), UC Berkeley Undergrad.
- Arnav Gudibande (2022–2023), UC Berkeley Masters. Published [8]. Now at Perplexity AI.
- Tony Zhao (2020–2021), UC Berkeley Undergrad. Published [18, 20]. Now PhD at Stanford.
- Albert Xu (2020–2021), UC Berkeley Undergrad. Published [15, 21]. Now PhD at USC.
- Eshaan Pathak (2020–2021), UC Berkeley Undergrad. Published [15, 21]. Now at You.com
- Jens Tuyls (2019–2020), UC Irvine Undergrad. Published [25, 29]. Now PhD at Princeton.
- Junlin Wang (2019–2020), UC Irvine Undergrad. Published [25, 29]. Now PhD at Duke.
- Nikhil Kandpal (2019), UMD Undergrad. Published [28]. Now PhD at UNC.

Masters Thesis Advising

- Arnav Gudibande, 2023. *On Imitating Proprietary Language Models.* Chair: Dawn Song.

Other Mentoring

- BAIR Undergraduate Mentoring, 2022–2024.
- Association of Women in EE&CS Office Hours on Grad School, 2023.
- 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

Presentations

External Talks

- ICLR Workshop on Secure and Trustworthy LLMs. *Making "GPT-Next" Trustworthy*
- ICLR Workshop on Data for Foundation Models. *Making "GPT-Next" Trustworthy*
- IEEE S&P Workshop on Security Architectures for GenAI. *Making "GPT-Next" Trustworthy*
- Simons Institute Workshop on LLMs, 2023. *Memorization in Large Language Models*
- Simons Institute Workshop on LLMs, 2023. *Memorization in Large Language Models*
- Princeton, 2023. *Memorization in Large Language Models*
- Oracle Labs, 2023. *Memorization in Large Language Models*
- University of Maryland, 2023. *Memorization in Large Language Models*
- University of North Carolina, 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?*

Panels:

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

Academic Service

Program Committee Member

- Journals: Computational Linguistics (2023)
- Conferences: ACL (2020, 2021, 2022), ICML (2021, 2023), NeurIPS (2020, 2021), EMNLP (2018, 2019, 2020, 2021, 2022), ACL Rolling Review (2021, 2022), ICLR (2023), NAACL (2021, 2022), COLM (2024)
- Workshops: Distribution Shifts (NeurIPS 2022, ICML 2022, NeurIPS 2023), BlackBox NLP (EMNLP 2022), RobustML Workshop (ICLR 2021), MRQA (EMNLP 2021), NLP for Positive Impact (ACL 2021), SRW (NAACL 2021), DistShift (NeurIPS 2021, NeurIPS 2023)

Departmental Service

- Berkeley PhD Admissions. 2021–2023
- Berkeley Student Committee for Faculty Hiring. 2023
- Berkeley PhD Visit Days Recruitment. 2021–2024

Academic Grants & Sponsorships

- Led successful award for ∼1500 TPUs from Google TRC ($10M+ USD value)
- Apple Fellowship in AI/ML, 2022–2024

Workshop Organization

- Future of Decentralization, AI, and Computing Summit. 2023, Berkeley. Led by Dawn Song.

Selected Media & Press

Extracting Training Data from Diffusion Models [9], MIT Technology Review, TWIML Podcast, Gizmodo, Vice, TechSpot, New Scientist, The Register, Ars Technica, Twitter #1 (3 million views), Twitter #2, Twitter #3,

Automated Crossword Solving [15], 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, Twitter (1M views)

Extracting Training Data from Large Language Models [19], MIT Technology Review, Wired, Google Blog, BAIR Blog, Nature, Top of Hacker News, Twitter #1, Twitter #2, Twitter #3,