# Sarah A. Wiegreffe

✉ *wiegreffesarah@gmail.com*
🖱 *sarahwie.github.io*

## Education

2022–present **Allen Institute for AI (Ai2)**, *Postdoctoral Researcher*.
Post-doctoral position ("young investigator") advised by Dr. Ashish Sabharwal and Professor Hannaneh Hajishirzi. Hold a courtesy appointment in the Paul G. Allen School of Computer Science and Engineering at the University of Washington.

2017–2022 **Georgia Institute of Technology**, *Ph.D. in Computer Science*.
Advisor: Professor Mark Riedl.
Dissertation: *Interpreting Neural Networks for and with Natural Language*.
Committee: Professors Alan Ritter, Wei Xu, Noah Smith (University of Washington), and Sameer Singh (University of California Irvine).

2017–2020 **Georgia Institute of Technology**, *M.S. in Computer Science*.
Specialization: Machine Learning.
Relevant coursework: Computational Statistics, Statistical Machine Learning, Deep Learning, Natural Language Processing.

2013–2017 **Honors College at the College of Charleston**, *B.S. in Data Science*.
**Summa Cum Laude.**
Awarded Data Science Major of the Year and Departmental Honors.
Minors in Mathematics and International Studies.

2015 **University of Tartu**, *Estonia*.
Visiting student in the Faculty of Mathematics and Computer Science.
Coursework: Cryptology, Computational Neuroscience, Advanced French (European scale B2→C1).

## Publications

Acceptance rates listed where known. * denotes equal contribution.

### In Submission

2025 Aaron Mueller*, Atticus Geiger*, **Sarah Wiegreffe***, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, Yonatan Belinkov. *The Mechanistic Interpretability Localization Benchmark.*

### Peer-reviewed, Archival

ICLR 2025 Jack Merullo, Noah A. Smith, **Sarah Wiegreffe*** & Yanai Elazar*. *On Linear Representations and Pretraining Data Frequency in Language Models.* International Conference on Learning Representations. Acceptance rate 32.08%. Also **one of 4 papers selected for oral presentation** at the ATTRIB workshop, NeurIPS 2024.

ICLR 2025 **Sarah Wiegreffe**, Oyvind Tafjord, Yonatan Belinkov, Hannaneh Hajishirzi, Ashish Sabharwal. *Answer, Assemble, Ace: Understanding How LMs Answer Multiple Choice Questions.* International Conference on Learning Representations. Acceptance rate 32.08%. **Spotlight (top 5.1% of submissions).**

NeurIPS 2024 Datasets & Benchmarks Faeze Brahman, Sachin Kumar, Vidhisha Balachandran* & Pradeep Dasigi* & Valentina Pyatkin* & Abhilasha Ravichander* & **Sarah Wiegreffe***, Nouha Dziri, Khyathi Chandu, Jack Hessel, Yulia Tsvetkov, Noah A. Smith, Yejin Choi, Hannaneh Hajishirzi. *The Art of Saying No: Contextual Noncompliance in Language Models.* Conference on Neural Information Processing Systems Datasets and Benchmarks Track. Acceptance rate 25.3%.

EMNLP 2024 BlackboxNLP Workshop Naomi Saphra* & **Sarah Wiegreffe***. *Mechanistic?* **One of 4 papers selected for oral presentation (top 6.25% of submissions).**

EMNLP 2024 Findings Shramay Palta, Nishant Balepur, Peter A. Rankel, **Sarah Wiegreffe**, Marine Carpuat, Rachel Rudinger. *Plausibly Problematic Questions in Multiple-Choice Benchmarks for Commonsense Reasoning.* Findings of the Conference on Empirical Methods in Natural Language Processing. Acceptance rate: 37.7%.

EMNLP 2024 Findings Yanai Elazar, Bhargavi Paranjape* & Hao Peng* & **Sarah Wiegreffe***, Khyathi Raghavi Chandu, Vivek Srikumar, Sameer Singh, Noah A. Smith. *Measuring and Improving Attentiveness to Partial Inputs with Counterfactuals.* Findings of the Conference on Empirical Methods in Natural Language Processing. Acceptance rate: 37.7%.

ACL 2024 Peter Hase, Mohit Bansal, Peter Clark, **Sarah Wiegreffe**. *The Unreasonable Effectiveness of Easy Training Data for Hard Tasks.* Annual Meeting of the Association for Computational Linguistics. **Led to invited talks at UC Berkeley and OpenAI.**

NeurIPS 2023 Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, **Sarah Wiegreffe**, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, Peter Clark. *Self-Refine: Iterative Refinement with Self-Feedback.* Conference on Neural Information Processing Systems. Acceptance rate 26.1%.

EMNLP 2023 **Sarah Wiegreffe**, Matthew Finlayson, Oyvind Tafjord, Peter Clark, Ashish Sabharwal. *Increasing Probability Mass on Answer Choices Does Not Always Improve Accuracy.* Conference on Empirical Methods in Natural Language Processing. Acceptance rate 21.3%.

EMNLP 2023 Anshita Gupta, Debanjan Mondal, Akshay Krishna Sheshadri, Wenlong Zhao, Xiang Lorraine Li* & **Sarah Wiegreffe*** & Niket Tandon*. *Editing Common Sense in Transformers.* Conference on Empirical Methods in Natural Language Processing. Acceptance rate 21.3%.

EMNLP 2022 Findings Kaige Xie, **Sarah Wiegreffe**, Mark Riedl. *Calibrating Trust of Multi-Hop Question Answering Systems with Decompositional Probes.* Findings of the Conference on Empirical Methods in Natural Language Processing. Acceptance rate 32.9%.

EMNLP 2022 Findings Xiangyu Peng, Siyan Li, **Sarah Wiegreffe**, Mark Riedl. *Inferring the Reader: Guiding Automated Story Generation with Commonsense Reasoning.* Findings of the Conference on Empirical Methods in Natural Language Processing. Acceptance rate 32.9%.

NAACL 2022 **Sarah Wiegreffe**, Jack Hessel, Swabha Swayamdipta, Mark Riedl, Yejin Choi. *Reframing Human-AI Collaboration for Generating Free-Text Explanations.* Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Acceptance rate 22.0%. **Led to invited talk at Oxford.**

NeurIPS 2021 Datasets & Benchmarks **Sarah Wiegreffe**$^*$ & Ana Marasović$^*$. *Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing.* Conference on Neural Information Processing Systems Datasets and Benchmarks Track. Acceptance rate 38%.

EMNLP 2021 **Sarah Wiegreffe**, Ana Marasović, Noah A. Smith. *Measuring Association Between Labels and Rationales.* Conference on Empirical Methods in Natural Language Processing. Acceptance rate 23.4%. **Led to invited talk at NLP with Friends.**

ACL 2020 Sarthak Jain, **Sarah Wiegreffe**, Yuval Pinter, Byron C. Wallace. *Learning to Faithfully Rationalize by Construction.* Annual Meeting of the Association for Computational Linguistics. Acceptance rate 22.7%.

EMNLP 2019 **Sarah Wiegreffe**$^*$ & Yuval Pinter$^*$. *Attention is not not Explanation.* Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing. Acceptance rate 24%. **Led to invited talks at USC and the "Big Picture" retrospective workshop at EMNLP 2023.**

ACL 2019 BioNLP Workshop **Sarah Wiegreffe**, Edward Choi, Sherry Yan, Jimeng Sun, Jacob Eisenstein. *Clinical Concept Extraction for Document-Level Coding.* Biomedical Natural Language Processing Workshop (BioNLP) at the Annual Meeting of the Association for Computational Linguistics.

NAACL 2018 James Mullenbach, **Sarah Wiegreffe**, Jon Duke, Jimeng Sun, Jacob Eisenstein. *Explainable Prediction of Medical Codes from Clinical Text.* Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Acceptance rate 31%.

### Peer-reviewed, Non-archival (poster presentations)

- Jack Merullo, Sarah Wiegreffe$^*$ & Yanai Elazar$^*$. *The Mutual Relationship between Corpus Frequency and Linear Representations in Language Models*. Talk, poster & non-archival paper at Workshop on Attributing Model Behavior at Scale (ATTRIB), NeurIPS 2024.
- Xiangyu Peng, Siyan Li, Sarah Wiegreffe, Mark Riedl. *Inferring the Reader: Guiding Automated Story Generation with Commonsense Reasoning*. Poster at Generation, Evaluation & Metrics (GEM) Workshop, EMNLP 2022.
- Kaige Xie, Sarah Wiegreffe, Mark Riedl. *Calibrating Trust of Multi-Hop Question Answering Systems with Decompositional Probes.* Poster at BlackBoxNLP Workshop, EMNLP 2022.
- Xiangyu Peng, Siyan Li, Sarah Wiegreffe, Mark Riedl. *Inferring the Reader: Guiding Automated Story Generation with Commonsense Reasoning.* Poster & non-archival paper at Narrative Understanding Workshop, NAACL 2021.

- Xiangyu Peng, Siyan Li, Sarah Wiegreffe, Mark Riedl. *Improving Neural Storytelling with Commonsense Inferences.* Poster & extended abstract at Women in Machine Learning (WiML) Workshop, NeurIPS 2020.
- Sarah Wiegreffe, Yuval Pinter. *Attention is not not Explanation.* Poster & extended abstract at Women in Machine Learning (WiML) Workshop, NeurIPS 2019.
- Sarah Wiegreffe, Jihad Obeid, Paul Anderson. *Can Classification of Publications by Translational Categories be Automated?* Poster & extended abstract at the American Medical Informatics Association (AMIA) Translational Bioinformatics Summit 2017.

## Selected Honors and Awards

2024 **Rising Star in Machine Learning**, *University of Maryland*.
One of 6 sponsored to attend a 2-day academic career workshop.

2024 **Outstanding Area Chair**, *Association for Computational Linguistics*.
Awarded to the top area chairs at the EMNLP 2024 conference.

2024 **Rising Star in Generative AI**, *University of Massachusetts, Amherst*.
Awarded to 9 people on the faculty market. Sponsored to attend a 2-day academic career workshop.

2023 **Top Reviewer**, *NeurIPS*.
Awarded to the top ~10% of reviewers. Granted free registration.

2023 **Rising Star in EECS**, *Georgia Institute of Technology*.
Acceptance rate 35% across all universities. Sponsored to attend a 2-day academic career workshop.

2023 **Outstanding Area Chair**, *Association for Computational Linguistics*.
Awarded to the top 1.5% of area chairs and reviewers at the ACL 2023 conference. Granted free virtual registration.

2020 **Outstanding Intern**, *Allen Institute for AI*.
Gift of $10,000 and returning offer. Awarded to 2-3 interns per year by research mentor nomination.

2018-2021 **Travel Awards**, *Various organizations*.
Received over $4000 outside of advisor funding to attend conferences over the course of my PhD.

2018 **Graduate Cohort Member**, *ACM Computing Research Association*.
Sponsored to attend the Association for Computing Machinery (ACM)'s national workshop for female computing PhD students.

2017 **Graduate Fellowship**, *Phi Kappa Phi Honor Society*.
Gift of $5,000. Awarded to 51 students nationwide beginning doctoral studies.

## Selected Invited Talks

2023 **Is "Attention = Explanation"? Past, Present, and Future**, *Keynote with Sarthak Jain at "The Big Picture" Workshop*, EMNLP 2023.

2023 **What is AI?**, *Committee on Environment, Energy, and Technology*, Washington State Senate.

2023 **Towards Transparent Language Models**, *Seminar talks at USC, UC Irvine, and UCSD*.

2023 **Two Views of Language Model Interpretability**, *Keynote at the Workshop on Natural Language Reasoning and Structured Explanations*, ACL 2023.

2022 **On Understanding and Explaining Large Language Models- what's missing?**, *Computational Linguistics Seminar*, University of Washington.

2022 **Reframing Human-AI Collaboration for Generating Free-Text Explanations**, *University of Oxford*.

2021 **Measuring Association Between Labels and Free-Text Rationales**, *NLP with Friends seminar (online)*.

2020 **BlackBoxNLP: What are we looking for, and where do we stand?**, *NLP/ISI seminar*, University of Southern California.

## Teaching

### Tutorials

NAACL 2024 **Explanation in the Era of Large Language Models**, *Expected attendance: 200*.

### Assistantships

Fall 2021 **Natural Language Processing (CS 7643)**, *Georgia Tech*, 91 students.

Spring 2021 **Deep Learning (CS 4803/7643)**, *Georgia Tech*, 170 students.

Fall 2019 **Deep Learning (CS 4803/7643)**, *Georgia Tech*, 215 students.
Pushed to include content on Transformers, gave the inaugural course lecture on the topic, and created an associated coding assignment from scratch. Student feedback was positive.

Spring 2019 **Machine Learning (CS 4641)**, *Georgia Tech*, 110 students.

### Guest Lectures

2024 **Towards Transparent Language Models**, Graduate Large Language Models course at Washington University in St. Louis.

2019 **Transformers and Natural Language Applications**, Graduate Deep Learning course at Georgia Tech.

### Advising & Mentoring (met at least weekly during course of project)

2024-present **Jack Merullo**, *PhD student at Brown University*.
Ai2 intern working on the relationship between pretraining data and linear structures in language model hidden states. Resulted in an ICLR paper and an oral workshop presentation.

2024-present **Alec Bunn**, *Undergraduate student at the University of Washington*.

2023-present **Shramay Palta**, *PhD student at the University of Maryland*.
Resulted in an EMNLP Findings paper and an ongoing followup project.

2023-2024 **Peter Hase**, *PhD student at UNC Chapel Hill*.
Ai2 intern working on methods for generating predictions from language models that generalize from easy to hard tasks when labeled data is scarce. Resulted in an ACL paper.

2023 **Joris Baan**, *PhD student at University of Amsterdam/ELLIS*.
Ai2 intern working on quantifying uncertainty in language models' textual generations.

2023 **Anshita Gupta, Debanjan Mondal, and Akshay Krishna Sheshadri**, *Master's students at UMass Amherst*.
Resulted in an EMNLP paper.

2021–2022 **Kaige Xie**, *Machine Learning PhD student at Georgia Tech*.
Resulted in an EMNLP Findings paper and a workshop presentation.

2020–2022 **Xiangyu Peng**, *Machine Learning PhD student at Georgia Tech*.
**and Siyan Li**, *Undergraduate student at Georgia Tech*.
Resulted in an EMNLP Findings paper and three workshop presentations.

## Academic Service

### Organization

- Area Chair: *EMNLP 2022, ACL 2023 (outstanding area chair), EMNLP 2023, ACL Rolling Review (2024- inc. EMNLP 2024 (outstanding area chair))*
- Workshop Organizer: *BlackBoxNLP 2022, 2 submissions to ICML 2025*
- Publicity Chair: *NAACL 2021*
- Birds-of-a-Feather Host: *NAACL 2021* (online), *NAACL 2022* (in person/hybrid)
- Student Volunteer: *EMNLP 2019, FAT* 2019, NAACL 2018*

### Conference/Journal Reviewing

- Computational Linguistics: *2025*
- ICLR: *2025*
- COLM: *2024*
- ICML: *2024*
- NeurIPS: *2023 (outstanding reviewer)*
- AI Magazine: *2023*
- Transactions on Interactive Intelligent Systems (TiiS): *2022, 2023*
- ACL Rolling Review (ARR): *Nov & Dec 2021; March & Oct 2022; Dec 2023*
- NAACL: *2021*
- EMNLP: *2019, 2020, 2021*
- ACL: *2018 (subreviewer), 2019, 2020*
- AMIA Informatics: *2018, 2019*

### Workshop Reviewing

- BlackBoxNLP (EMNLP): *2020, 2021, 2023*
- Deep Learning Approaches for Low-Resource NLP (NAACL): *2022*
- Commonsense Representation and Reasoning (ACL): *2022*
- Women in Machine Learning (NeurIPS): *2019*
- Machine Learning for Healthcare (NeurIPS): *2017, 2018, 2019*

### Outreach

- Consulting to staffers in U.S. Senate Chamber of Commerce about explainable AI: *2024*
- “What is AI?”, talk given to the Washington State Senate: *2023*
- Reviewer, Georgia Tech PhD Application Support Program for underrepresented applicants: *2021*

- Panelist, College of Charleston Honors College "How to Tell If (and When) Graduate School is Right for You": *2020*

## Professional Experience

### Industry

2021 **Research Intern**, *Allen Institute for AI*.
Hosted by Drs. Jack Hessel and Swabha Swayamdipta, and Professor Yejin Choi. Worked on few-shot explanation generation and effective human evaluation.

2020 **Research Intern**, *Allen Institute for AI*.
Hosted by Dr. Ana Marasović and Professor Noah Smith. Worked on interpretability of deep learning models for NLP. **Awarded outstanding intern award.**

2019 **Research Intern**, *Google AI Health (formerly/now Google Brain/Deepmind)*.
Hosted by Dr. Edward Choi (now assistant professor at KAIST), Gerardo Flores, and Dr. Andrew Dai. Improved outcome prediction for clinical time-series data using unsupervised pretraining. Resulted in unpublished short paper *Learning Bi-Directional Clinical Event Representations: a Comparison of Architectures* (available upon request).

2018 **Research Intern**, *Sutter Health*.
Hosted by Dr. Sherry Yan and Professor Jimeng Sun. Worked on deep learning methodology for disease prediction from clinical text.

## Press

2023 **The frightening truth about AI chatbots: Nobody knows exactly how they work**, *Fast Company*.