Atticus Geiger

Stanford University
atticusg@stanford.edu

Education

Ph.D. Linguistics *September 2019 - December 2023*
Stanford University, Stanford, CA

M.S. Computer Science *September 2016 - June 2019*
Stanford University, Stanford, CA

B.S. Symbolic Systems *September 2015- June 2019*
Stanford University, Stanford, CA

Career History

Member of Technical Staff, *Goodfire* *September 2025 -*
Principal Investigator, *Practical AI Alignment and Interpretability Research Group* *October 2023 -*

- Acquired grant from Open Philanthropy to start a remote, non-profit research group focused on analyzing and interpreting AI models
- Support students internationally with remote research internship opportunities
- Conduct interpretability research grounded in theories of causality and abstraction

PhD Research, *Stanford University* *September 2019 - December 2023*

- Advised by Christopher Potts and Thomas Icard (Also mentored by Noah Goodman and Mike Frank)
- Primary Author Publications at Psychological Review, NeurIPS, ICML, EMNLP, ACL, NAACL, CLeaR, CogSci and BlackBoxNLP
- Collaborated with and mentored several graduate students and undergraduates

Honors Thesis Research, *Stanford University* *June 2018 - September 2018*

- Acquired grant for self led natural language inference research project advised by Chris Potts, Thomas Icard, and Lauri Karttunen
- Constructed artificial natural language inference dataset using logic models
- Designed task specific neural model with standout performance on the generated datasets

Symbolic Systems Research Intern, *Stanford University* *June 2017 - September 2017*

- Worked with Lauri Karttunen and Ignacio Cases to create a Natural Language Inference dataset focused on implicatives
- Implemented neural network models for natural language inference in Tensor Flow

Software Engineer Intern, *Alaska Satellite Facility* *June 2016 - September 2016*

- Created a AWS cloud processing infrastructure and website interface
- Provided a service that automatically processes incoming satellite data
- Accomplished significant, largely self led software implementation

Skills

Programming: Python, C++, C, TensorFlow, PyTorch, AWS
Social Skills: I love collaborating and find great joy in mentoring!

Awards and Honors

Firestone Medal Award for B.S. honors thesis titled *Can Natural Language Inference Models Perform Natural Logic Reasoning?* and advised by Chris Potts and Thomas Icard.

Grants

PI, Fair Adversarial Tasks for Natural Language Understanding. Facebook Research: Robust Deep Learning for Natural Language Processing. Co-PI Christopher Potts. 2019-2020
PI, Practical AI Alignment and Interpretability Research Group. Open Philanthropy. 2023-2025

Publications

Primary author: EMNLP, NAACL, NeurIPS (x2), ICML (x3), CogSci, JMLR, Psychological Review, BlackBoxNLP, and CLeaR
Senior author: ICLR, ACL (x2), EMNLP, BlackBoxNLP, and CLeaR
Middle author: NeurIPS (x2), ICML (x2), NAACL (x3), ACL, CLeaR, and BlackBoxNLP (x2)

1. Or Shafran, **Atticus Geiger**, and Mor Geva. Decomposing mlp activations into interpretable features via semi-nonnegative matrix factorization, 2025
2. Nikhil Prakash, Natalie Shapira, Arnab Sen Sharma, Christoph Riedl, Yonatan Belinkov, Tamar Rott Shaham, David Bau, and **Atticus Geiger**. Language models use lookbacks to track beliefs, 2025
3. Jiuding Sun, Sidharth Baskaran, Zhengxuan Wu, Michael Sklar, Christopher Potts, and Atticus Geiger. Hypersteer: Activation steering at scale with hypernetworks, 2025
4. Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, **Atticus Geiger**, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. Open problems in mechanistic interpretability, 2025
5. Yiwei Wu, **Atticus Geiger**, and Raphaël Millière. How do transformers learn variable binding in symbolic programs? In *Forty-second International Conference on Machine Learning*, 2025
6. 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, and Yonatan Belinkov. MIB: A mechanistic interpretability benchmark. In *Forty-second International Conference on Machine Learning*, 2025
7. Zhengxuan Wu, Aryaman Arora, **Atticus Geiger**, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Axbench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025
8. Jiuding Sun, Jing Huang, Sidharth Baskaran, Karel D'Oosterlinck, Christopher Potts, Michael **Sklar***, and **Atticus Geiger***. HyperDAS: Towards automating mechanistic interpretability with hypernetworks. In *The Thirteenth International Conference on Learning Representations*, 2025
9. **Atticus Geiger**, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025
10. Yoav Gur-Arieh, Roy Mayan, Chen Agassy, **Atticus Geiger**, and Mor Geva. Enhancing automated interpretability with output–centric feature descriptions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*, 2025. To appear
11. Theodora-Mara Pîslar, Sara Magliacane, and Atticus Geiger. Combining causal models for more accurate abstractions of neural networks. In Biwei Huang and Mathias Drton, editors, *Proceedings of the Fourth Conference on Causal Learning and Reasoning*, volume 275 of *Proceedings of Machine Learning Research*, pages 114–138. PMLR, 07–09 May 2025
12. Róbert Csordás, Christopher Potts, Christopher D Manning, and **Atticus Geiger**. Recurrent neural networks learn to store and generate sequences using non-linear representations. In Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Aaron Mueller, and Hanjie Chen, editors, *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 248–262, Miami, Florida, US, November 2024. Association for Computational Linguistics
13. Zhengxuan Wu, Aryaman Arora, Zheng Wang, **Atticus Geiger**, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. Reft: Representation finetuning for language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024

14. Curt Tigges, Oskar J. Hollinsworth, **Atticus Geiger**, and Neel Nanda. Language models linearly represent sentiment. In Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Aaron Mueller, and Hanjie Chen, editors, *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 58–87, Miami, Florida, US, November 2024. Association for Computational Linguistics

15. Amir Zur, Elisa Kreiss, Karel D'Oosterlinck, Christopher Potts, and **Atticus Geiger**. Updating CLIP to prefer descriptions over captions. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20178–20187, Miami, Florida, USA, November 2024. Association for Computational Linguistics

16. **Atticus Geiger***, Zhengxuan Wu*, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In Francesco Locatello and Vanessa Didelez, editors, *Causal Learning and Reasoning, 1-3 April 2024, Los Angeles, California, USA*, volume 236 of *Proceedings of Machine Learning Research*, pages 160–187. PMLR, 2024

17. Zhengxuan Wu, **Atticus Geiger**, Aryaman Arora, Jing Huang, Zheng Wang, Noah D. Goodman, Christopher D. Manning, and Christopher Potts. pyvene: A library for understanding and improving pytorch models via interventions. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2024)*, Mexico City, Mexico, June 2024. Association for Computational Linguistics. Demonstrations Track

18. Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and **Atticus Geiger**. RAVEL: evaluating interpretability methods on disentangling language model representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, Bangkok, Thailand, August 2024. Association for Computational Linguistics

19. Zhengxuan Wu, **Atticus Geiger**, Jing Huang, Aryaman Arora, Thomas Icard, Christopher Potts, and Noah D. Goodman. A reply to Makelov et al. (2023)'s "interpretability illusion" arguments. arXiv:2401.12631, 2024

20. Zhengxuan Wu*, **Atticus Geiger***, Christopher Potts, and Noah D. Goodman. Interpretability at scale: Identifying causal mechanisms in Alpaca. In *Advances in Neural Information Processing Systems*, 2023

21. Jing Huang, **Atticus Geiger**, Karel D'Oosterlinck, Zhengxuan Wu, and Christopher Potts. Rigorously assessing natural language explanations of neurons. In *Proceedings of the Sixth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, December 2023

22. Jingyuan S. She, Christopher Potts, Samuel R. Bowman, and **Atticus Geiger**. ScoNe: Benchmarking negation reasoning in language models with fine-tuning and in-context learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1803–1821, Toronto, Canada, July 2023. Association for Computational Linguistics

23. Angela Cao*, **Atticus Geiger***, Elisa Kreiss*, Thomas Icard, and Tobias Gerstenberg. A semantics for causing, enabling, and preventing verbs using structural causal models. In *Proceedings of the Cognitive Science Society*, 2023

24. Riccardo Massidda, **Atticus Geiger**, Thomas Icard, and Davide Bacciu. Causal abstraction with soft interventions. In *2nd conference on Causal Learning and Reasoning (CLeaR)*, Tbingen, Germany, 2023

25. **Atticus Geiger**, Alexandra Carstensen, Michael C Frank, and Christopher Potts. Relational reasoning and generalization using nonsymbolic neural networks. *Psychol. Rev.*, 130(2):308–333, March 2023

26. Zhengxuan Wu*, Karel D'Oosterlinck*, **Atticus Geiger***, Amir Zur, and Christopher Potts. Causal Proxy Models for concept-based model explanations. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. arXiv:2209.14279

27. Eldar David Abraham*and Karel D'Oosterlinck*and Amir Feder*and Yair Ori Gat*and **Atticus Geiger***and Christopher Potts*and Roi Reichart*and Zhengxuan Wu*. Cebab: Estimating the causal effects of real-world concepts on NLP model behavior. *CoRR*, abs/2205.14140, 2022

28. Zhengxuan Wu*, **Atticus Geiger***, Josh Rozner, Elisa Kreiss, Hanson Lu, Thomas Icard, Christopher Potts, and Noah D. Goodman. Causal distillation for language models. In *Proceedings of the 2022*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4288–4295, Seattle, United States, July 2022. Association for Computational Linguistics

29. **Atticus Geiger***, Zhengxuan Wu*, Hanson Lu*, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 7324–7338. PMLR, 17–23 Jul 2022
30. **Atticus Geiger***, Hanson Lu*, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, 2021
31. Christopher Potts, Zhengxuan Wu, **Atticus Geiger**, and Douwe Kiela. DynaSent: A dynamic benchmark for sentiment analysis. In *Proceedings of the Association for Computational Linguistics*, 2021
32. Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, **Atticus Geiger**, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online, June 2021. Association for Computational Linguistics
33. **Atticus Geiger**, Kyle Richardson, and Christopher Potts. Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 163–173, Online, November 2020. Association for Computational Linguistics
34. **Atticus Geiger**, Ignacio Cases, Lauri Karttunen, and Christopher Potts. Posing fair generalization tasks for natural language inference. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4475–4485, Stroudsburg, PA, November 2019. Association for Computational Linguistics
35. **Atticus Geiger**. Can natural language inference models perform natural logic reasoning? B.s. thesis, Stanford University, 2019
36. Ignacio Cases, Clemens Rosenbaum, Matthew Riemer, **Atticus Geiger**, Tim Klinger, Alex Tamkin, Olivia Li, Sandhini Agarwal, Joshua D. Greene, Dan Jurafsky, Christopher Potts, and Lauri Karttunen. Recursive routing networks: Learning to compose modules for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, June 2019. Association for Computational Linguistics
37. **Atticus Geiger**, Ignacio Cases, Lauri Karttunen, and Christopher Potts. Stress-testing neural models of natural language inference with multiply-quantified sentences. Ms., Stanford University. arXiv 1810.13033, 2018
38. Steve A. Arko, Rose Hogenson, **Atticus Geiger**, Jake Herrmann, Brian Buechler, and Kirk Hogenson. Sentinel-1 Archive and Processing in the Cloud using the Hybrid Pluggable Processing Pipeline (HyP3) at the ASF DAAC. In *AGU Fall Meeting Abstracts*, volume 2016, pages G43A–1040, December 2016
39. Kirk. Hogenson, Steve A. Arko, Brian Buechler, Rose Hogenson, Jake Herrmann, and **Atticus Geiger**. Hybrid Pluggable Processing Pipeline (HyP3): A cloud-based infrastructure for generic processing of SAR data. In *AGU Fall Meeting Abstracts*, volume 2016, pages IN21B–1740, December 2016

Invited Talks

1. 2025. *How Do Transformers Learn Variable Binding in Symbolic Programs?*. Weekly Seminar: Deep Learning: Classics and Trends. August 25.
2. 2025. *Causal Abstraction as a Theory of Computational Implementation.* XAI Seminar @ Imperial College London. July 12.
3. 2025. *Benchmarking Methods for Understanding and Controlling Large Language Models.* 9th Annual CHAI workshop @ Asilomar. June 5-8

4. 2025. *Benchmarking Methods for Understanding and Controlling Large Language Models.* FAR AI Seminar. May 21
5. 2025. *Benchmarking Methods for Understanding and Controlling Large Language Models.* Apple Machine Learning Research Weekly Seminar. April 17.
6. 2025. *The Current State of Interpretability and Ideas for Scaling Up.* Theory of Interpretable AI Seminar. March 12.
7. 2025. *Generalizing Causal Abstraction to Intervention Algebras.* Bellairs Fourth Annual Workshop on Causality. February 27.
8. 2025. *Causal Abstraction and Causal Emergence in Mechanistic Interpretability.* Bellairs Fourth Annual Workshop on Causality. February 29.
9. 2024. *The Current State of Interpretability and Ideas for Scaling Up.* The Buzz Robot AI Community. December 19.
10. 2024. *The Current State of Interpretability and Ideas for Scaling Up.* Model Interventions (MINT) workshop at NeurIPS. December 15.
11. 2024. *Uncovering and Inducing Interpretable Causal Structure in Neural Networks.* Causality, Abstraction, Reasoning, and Extrapolation (CARE) Seminar. April 11.
12. 2024. *Causal Abstraction as a Theoretical Foundation for Mechanistic Interpretability.* Workshop on Logic and AI at The Institute for Advanced Study. July 17.
13. 2023. *Theories and Tools for Mechanistic Interpretability via Causal Abstraction.* AI Alignment Workshop at NeurIPS in New Orleans. December.
14. 2023. *Uncovering and Inducing Causal Structure in Deep Learning Models.* Symbolic Systems Weekly Seminar at Stanford. December 4.
15. 2023. *Theories and Tools for Mechanistic Interpretability via Causal Abstraction.* AI safety and mechanistic interpretability conference at MIT. May 6-7.
16. 2023. *Finding Alignments Between Interpretable Causal Variables and Distributed Neural Representations.* Workshop on Causal Representation Learning (CRL 2023) at the Max Planck Institute for Intelligent Systems in Tbingen. April 17-19th.
17. 2022. *Causal Abstraction and Computational Explanations in Artificial Intelligence.* 8th CSLI Workshop on Logic, Rationality, and Intelligent Interaction. May 22.
18. 2022. *Causal Abstraction and Computational Explanations in Artificial Intelligence.* Seminar Series in Stanford Psychology for cognitive and neuroscience areas. January 21.
19. 2021. *Causal Abstractions and Interchange Intervention Training.* Allen AI Aristo Group. September 23.
20. 2021. *Causal Abstraction for Neural Network Analysis.* McDonnell Plenary. July 8.
21. 2020. *Modular Representation in Neural Natural Language Inference Models.* Allen AI Aristo Group. June 19.

Teaching

1. 2025 (Fall). A modern explainable AI approach to theoretical neuroscience. Guest Lecture for AP 293. (Graduate) *Psycholinguistics* at Stanford. (Graduate and Undergraduate)
2. 2021 (Fall). Probabalistic Pragmatics. Guest Lecture for LINGUIST 245 *Psycholinguistics* at Stanford. (Graduate and Undergraduate)
3. 2020 (Spring). Causal Abstraction with Applications to Computational Implementation and Neural Network Analysis. Guest Lecture for PHIL 359 *Logic Seminar* at Stanford. (Graduate)
4. 2020 (Spring). Natural Logic. Guest Lecture for LINGUIST 130B *Lexical Semantics* at Stanford. (Undegraduate)
5. 2019 (Spring). Evaluating NLU Models with Harder Generalization Tasks. Guest Lecture for CS 224U *Natural Language Understanding* at Stanford. (Graduate)
6. Teaching Assistant Positions: CS 224U *Natural Language Understanding* at Stanford 2019 (Spring), LINGUIST 130B *Lexical Semantics* at Stanford 2020 (Spring) , and LINGUIST 245 *Psycholinguistics* at Stanford 2021 (Fall).