AARON MUELLER

CONTACT

Khoury College of Computer Sciences
Northeastern University
177 Huntington Ave., 22 Fl.
Boston, MA 02115 (USA)

aa.mueller@northeastern.edu
aaronmueller.github.io
github.com/aaronmueller

RESEARCH INTERESTS

- Natural language processing
- Mechanistic interpretability
- Computational psycholinguistics

EDUCATION

Johns Hopkins University — Baltimore, MD
Ph.D., Computer Science — Aug. 2023
M.S.E., Computer Science — May 2020
GPA: 3.9/4.0
Thesis title: Emergent Syntactic Behaviors and Mechanisms in Neural Language Models
Advisors: Tal Linzen, Mark Dredze
Committee: Tal Linzen, Mark Dredze, David Yarowsky, Yonatan Belinkov

New York University — New York, NY
Visiting academic, Center for Data Science — Aug. 2021 – Aug. 2023

University of Kentucky — Lexington, KY
B.S., Computer Science. *Honors* — May 2018
B.S., Linguistics. *Honors* — May 2018
GPA: 4.0/4.0. *Summa cum laude*

ACADEMIC POSITIONS

Northeastern University — Boston, MA
Zuckerman Postdoctoral Fellow, Khoury College of Computer Sciences — Aug. 2023 – Present
Advisor: David Bau

Technion – Israel Institute of Technology — Haifa, Israel
Zuckerman Postdoctoral Fellow, Department of Computer Science — Aug. 2023 – Present
Advisor: Yonatan Belinkov

INDUSTRY EXPERIENCE

Meta — Menlo Park, CA
Research Intern — May – Nov. 2022
Manager: Kanika Narang
- Research in retrieval-augmented generative models for few-shot question answering.
- Resulted in improved F_1 on multiple QA and classification datasets using far fewer parameters than state-of-the-art models. Also resulted in a publication at ACL [10].

Amazon Web Services (AWS) — Santa Clara, CA
Applied Scientist Intern — May – Aug. 2021
Manager: Saab Mansour
- Research in pre-training methods for improving goal-oriented dialogue agents.
- Resulted in state-of-the-art few-shot intent classification accuracy (>30% 1-shot gains) and a publication at ACL [15].

Raytheon BBN Technologies — Cambridge, MA
Research Intern — May – Aug. 2019
Manager: Ilana Heintz
- Research in low-resource cross-lingual word alignment and entity linking.
- Implemented convolutional neural machine translation models rivaling our prior seq2seq model's BLEU with over 20% faster training and over 50% faster inference.

PUBLICATIONS

Peer-reviewed Conference Articles

1. Michael Hanna*, **Aaron Mueller***. "Incremental Sentence Processing Mechanisms in Autoregressive Transformer Language Models." To appear in *Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, 2025. [*Equal contribution]
2. Jannik Brinkmann, Chris Wendler, Christian Bartelt, **Aaron Mueller**. "Large Language Models Share Representations of Latent Grammatical Concepts Across Typologically Diverse Languages." To appear in *Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, 2025.
3. Juan Diego Rodriguez, **Aaron Mueller**, Kanishka Misra. "Characterizing the Role of Similarity in the Property Inferences of Language Models." To appear in *Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, 2025.
4. Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, **Aaron Mueller**. "Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models." To appear in *International Conference on Learning Representations (ICLR)*, 2025.
5. Jaden Fiotto-Kaufman, Alexander R. Loftus, Eric Todd, Jannik Brinkmann, Caden Juang, Koyena Pal, Can Rager, **Aaron Mueller**, Samuel Marks, Arnab Sen Sharma, Francesca Lucchetti, Michael Ripa, Adam Belfki, Nikhil Prakash, Sumeet Multani, Carla Brodley, Arjun Guha, Jonathan Bell, Byron Wallace, David Bau. "NNsight and NDIF: Democratizing Access to Foundation Model Internals." To appear in *International Conference on Learning Representations (ICLR)*, 2025.
6. Yaniv Nikankin, Anja Reusch, **Aaron Mueller**, Yonatan Belinkov. "Arithmetic Without Algorithms: Language Models Solve Math With a Bag of Heuristics." To appear in *International Conference on Learning Representations (ICLR)*, 2025.
7. **Aaron Mueller**, Albert Webson, Jackson Petty, Tal Linzen. "In-context Learning Generalizes, But Not Always Robustly: The Case of Syntax." *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.
8. Eric Todd, Millicent L. Li, Arnab Sen Sharma, **Aaron Mueller**, Byron C. Wallace, David Bau. "Function Vectors in Large Language Models." In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.
9. **Aaron Mueller**, Tal Linzen. "How to Plant Trees in Language Models: Data and Architectural Effects on the Emergence of Syntactic Inductive Biases." In *Proceedings of the Association for Computational Linguistics (ACL)*, 2023.
10. **Aaron Mueller**, Kanika Narang, Lambert Mathias, Qifan Wang, Hamed Firooz. "Meta-training with Demonstration Retrieval for Efficient Few-shot Learning." In *Findings of the Association for Computational Linguistics (ACL)*, 2023.
11. Alex Warstadt*, **Aaron Mueller***, Leshem Choshen, Ethan Wilcox, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, Ryan Cotterell. "Findings of the BabyLM Challenge: Sample-efficient Pretraining on Developmentally Plausible Corpora." In *Proceedings of the BabyLM Challenge at the Conference on Computational Natural Language Learning (CoNLL)*, 2023. [*Equal contribution]
12. Koustuv Sinha, Jon Gauthier, **Aaron Mueller**, Kanishka Misra, Keren Fuentes, Roger Levy, Adina Williams. "Language Model Acceptability Judgements Are Not Always Robust to Context." In *Proceedings of the Association for Computational Linguistics (ACL)*, 2023. **Outstanding Paper Award.**
13. Julian Michael, Ari Holtzman, Alicia Parrish, **Aaron Mueller**, Alex Wang, Angelica Chen, Divyam Madaan, Nikita Nangia, Richard Yuanzhe Pang, Jason Phang, Samuel R. Bowman. "What Do NLP Researchers Believe? Results of the NLP Community Metasurvey." In *Proceedings of*

the Association for Computational Linguistics (ACL), 2023.

14. **Aaron Mueller**, Robert Frank, Tal Linzen, Luheng Wang, Sebastian Schuster. "Coloring the Blank Slate: Pre-training Imparts a Hierarchical Inductive Bias to Sequence-to-sequence Models." In *Findings of the Association for Computational Linguistics (ACL)*, 2022.
15. **Aaron Mueller**, Jason Krone, Salvatore Romeo, Saab Mansour, Elman Mansimov, Yi Zhang, Dan Roth. "Label Semantic Aware Pre-training for Few-shot Text Classification." In *Proceedings of the Association for Computational Linguistics (ACL)*, 2022.
16. **Aaron Mueller**, Yu Xia, Tal Linzen. "Causal Analysis of Syntactic Agreement Neurons in Multilingual Language Models." In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, 2022.
17. Alexandra DeLucia, Shijie Wu, **Aaron Mueller**, Carlos Aguirre, Mark Dredze, Philip Resnik. "BERNICE: A Multilingual Pre-trained Encoder for Twitter." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.
18. **Aaron Mueller**, Mark Dredze. "Fine-tuning Encoders for Improved Monolingual and Zero-shot Polylingual Neural Topic Modeling." In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021.
19. **Aaron Mueller**, Zach Wood-Doughty, Silvio Amir, Mark Dredze, Alicia L. Nobles. "Demographic Representation and Collective Storytelling in the Me Too Twitter Hashtag Activism Movement." In *Proceedings of the Association for Computing Machinery (ACM) on Human-Computer Interaction (HCI), Computer-Supported Cooperative Work and Social Computing (CSCW)*, 2021.
20. Matthew Finlayson*, **Aaron Mueller***, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, Yonatan Belinkov. "Causal Analysis of Syntactic Agreement Mechanisms in Neural Language Models." In *Proceedings of the Association for Computational Linguistics (ACL)*, 2021. [*Equal contribution]
21. **Aaron Mueller**, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, Tal Linzen. "Cross-linguistic Syntactic Evaluation of Word Prediction Models." In *Proceedings of the Association for Computational Linguistics (ACL)*, 2020.
22. **Aaron Mueller**, Garrett Nicolai, Arya D. McCarthy, Dylan Lewis, Winston Wu, David Yarowsky. "An Analysis of Massively Multilingual Neural Machine Translation for Low-Resource Languages." In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2020.
23. Arya D. McCarthy, Rachel Wicks, Dylan Lewis, **Aaron Mueller**, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, David Yarowsky. "The Johns Hopkins University Bible Corpus: 1600+ Tongues for Typological Exploration." In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2020.
24. Garrett Nicolai, Dylan Lewis, Arya D. McCarthy, **Aaron Mueller**, Winston Wu, David Yarowsky. "Fine-grained Morphosyntactic Analysis and Generation Tools for More Than One Thousand Languages." In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2020.
25. Marten van Schijndel, **Aaron Mueller**, Tal Linzen. "Quantity Doesn't Buy Quality Syntax with Neural Language Models." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.
26. Arya D. McCarthy, Winston Wu, **Aaron Mueller**, Bill Watson, David Yarowsky. "Modeling Color Terminology Across Thousands of Languages." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Peer-reviewed Journal Articles

1. Ethan Gotlieb Wilcox, Michael Hu, **Aaron Mueller**, Tal Linzen, Alex Warstadt, Leshem Choshen,

Chengxu Zhuang, Ryan Cotterell, Adina Williams. "Bigger Is Not Always Better: The Importance of Human-scale Language Modeling for Psycholinguistics." Accepted to *Journal of Memory and Language (JML)*.

2. Ian R. McKenzie, Alexander Lyzhov, Michael Martin Pieler, Alicia Parrish, **Aaron Mueller**, Ameya Prabhu, Euan McLean, Xudong Shen, Joe Cavanagh, Andrew George Gritsevskiy, Derik Kauffman, Aaron T. Kirtland, Zhengping Zhou, Yuhui Zhang, Sicong Huang, Daniel Wurgaft, Max Weiss, Alexis Ross, Gabriel Recchia, Alisa Liu, Jiacheng Liu, Tom Tseng, Tomasz Korbak, Najoung Kim, Samuel R. Bowman, Ethan Perez. "Inverse Scaling: When Bigger Isn't Better." In *Transactions on Machine Learning Research (TMLR)*, 2023. **Featured Paper.**

Peer-reviewed Workshop Articles

1. **Aaron Mueller**. "Missed Causes and Ambiguous Effects: Counterfactuals Pose Challenges for Interpreting Neural Networks." *The Mechanistic Interpretability Workshop at the International Conference on Machine Learning (ICML)*, 2024. **Honorable Mention for Top Paper.**
2. Aruna Sankaranarayanan, Dylan Hadfield-Menell, **Aaron Mueller**. "Disjoint Processing Mechanisms of Hierarchical and Linear Grammars in Large Language Models." In *ICML Workshop on Large Language Models and Cognition*, 2024.
3. Alexandra DeLucia*, **Aaron Mueller***, Xiang Lisa Li, João Sedoc. "Decoding Methods for Neural Narrative Generation." In *Proceedings of the Workshop on Generation Evaluation and Metrics (GEM) at Association for Computational Linguistics (ACL)*, 2021. [*Equal contribution]
4. **Aaron Mueller***, Yash Kumar Lal*. "Sentence-Level Adaptation for Low-Resource Neural Machine Translation." In *Proceedings of the Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT) at Machine Translation Summit (MTSummit)*, 2019. [*Equal contribution]

Preprints & In Submission

1. **Aaron Mueller**, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, Yonatan Belinkov. "The Quest for the Right Mediator: A History, Survey, and Theoretical Grounding of Causal Interpretability." arXiv preprint, 2024.

EDITOR

Workshop & Shared Task Proceedings

1. Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, **Aaron Mueller**, Hanjie Chen. "Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP." 2024.
2. Alex Warstadt, **Aaron Mueller**, Leshem Choshen, Ethan Wilcox, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, Ryan Cotterell. "Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning." 2023.

INVITED TALKS

Causal Abstraction and Causal Emergence in Mechanistic Interpretability.
Bellairs Workshop on Causality (Barbados). Feb. 19, 2025.

Understanding and Controlling Language Models at the Feature Level.
Geva Lab, Tel Aviv University (Tel Aviv, Israel). Dec. 18, 2024.

Interpretability for Computational Linguistics: Understanding and Modifying How Models Make Decisions.
Language, Computation, and Cognition Lab, The Technion (Haifa, Israel). Nov. 24, 2024.

Mechanistically Controlling Language Models.
- Department of Computer Science, École polytechnique fédérale de Lausanne (EPFL; Lausanne, Switzerland). Jul. 4, 2024.
- Department of Computer Science, Saarland University (Saarbrücken, Germany). Jul. 2, 2024.

Bigger Is Not Always Better: The Benefits of Building and Understanding Sample-efficient Language Models.
Department of Computer Science, Maastricht University (Maastricht, Netherlands). Jun. 26, 2024.

Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models.
NLP Seminar, University of California, Santa Barbara (Santa Barbara, CA). Apr. 24, 2024.

Evaluating and Surgically Improving Generalization in Language Models.
- Responsible AI Seminar Series, Nokia Bell Labs (Cambridge, UK). Mar. 18, 2024.
- NLP Seminar, University of Pittsburgh (Pittsburgh, PA). Feb. 29, 2024.
- Deep Learning Superlab, Brown University (Providence, RI). Feb. 15, 2024.

Planting Trees in Language Models: Emergent Syntactic Behaviors and Mechanisms from Pre-training.
- Koller Lab, Saarland University (Saarbrücken, Germany). Feb. 7, 2023.
- NLP Seminar, Technion – Israel Institute of Technology (Haifa, Israel). Dec. 14, 2022.
- Bar-Ilan NLP Seminar, Bar-Ilan University (Ramat Gan, Israel). Dec. 13, 2022.

What Generalizations do Sequence-to-sequence Models Learn from Multilingual Text? Insights from Translation and Syntactic Transformations.
Multilingual Text Processing Group, National Research Council of Canada (Ottawa, ON). Mar. 4, 2022.

Syntactic Agreement in Neural Language Models: How Well and Where Do They Perform Subject-Verb Agreement?
Language & Understanding Group, Mila – Québec Artificial Intelligence Institute (Montréal, QC). Mar. 22, 2021.

Causal Mediation Analysis for Analyzing Neural Networks.
Fairness & Interpretability Research Talk Series, Google (New York, NY). Mar. 17, 2021.

Causal Analysis of Syntactic Agreement Mechanisms in Neural Language Models.
Center for Language & Speech Processing Seminar, Johns Hopkins University (Baltimore, MD). Feb. 12, 2021.

FELLOWSHIPS AND AWARDS

Zuckerman Fellow, *International* 2023–2025
Two-year postdoctoral fellowship. Supports research joint with an Israeli university and an American university. ($126,000)

Microsoft Accelerate Foundation Models Research Award, *International* 2023
Awarded for research on the capabilities of large language models. Provides OpenAI API credits and priority GPT-4 access. ($10,000)

National Science Foundation Graduate Research Fellow, *National* 2018 - 2023
Five-year graduate research fellowship. Provides three years of Ph.D. funding. ($135,000)

Gaines Fellow, *University of Kentucky* 2016 - 2018
Two-year fellowship. Requires the completion of a juried project, a thesis project, and a seminar in the humanities. ($5,000)

Patterson Scholar, *University of Kentucky* 2014 - 2018
Four-year scholarship covering tuition, educational materials, and room & board. Awarded to undergraduates who have earned National Merit semifinalist standing or higher. ($86,000)

Goldwater Scholarship (Honorable Mention), *National* 2017

Phi Beta Kappa, *National* 2017

Raymond F. Betts Scholar, *University of Kentucky* 2017
Awarded for thesis research. Used funds to design language technologies for low-resource dialects of French. ($2,500)

Linguistics Research Award, *University of Kentucky* 2016
Awarded to an undergraduate to facilitate a year-long research project in linguistics. ($500)

MENTORING

Ph.D. students

- Aruna Sankaranarayanan (MIT). Joint with Dylan Hadfield-Menell. Research on natural and artificial grammar learning in language models. 2023–2025
- Juan Diego Rodriguez (UT Austin). Joint with Kanishka Misra. Research in how concepts are organized in language models. 2023–2024
- Eric Todd (Northeastern). Joint with David Bau. Research on how functions are represented in neural language models. Resulted in a publication at ICLR [8]. 2023–2024

Master's students

- Tal Haklay (Technion). Research on mechanistically understanding gender bias in language models. 2024–2025
- Dan Pechi (NYU). Research on imparting better inductive biases to language models. 2023
- Swapnil Sharma (NYU). Research on evaluating summarization models. 2022–2023
- Yash Kumar Lal (JHU). Resulted in a workshop publication [4]. 2018–2019

Undergraduate researchers

- Yu Xia (NYU). Resulted in a publication at CoNLL [16]. 2021–2022
- Matthew Finlayson (Harvard). Resulted in a publication at ACL [20]. 2020–2021

BU qualifying exam committees

- Gabriel Franco (Computer Science) 2025

TEACHING & LECTURES

Massachusetts Institute of Technology
Guest Lecture
Instructor: Jacob Andreas

- Quantitative Methods for NLP Fall 2024

New York University
Guest Lecture
Instructor: Tal Linzen

- Computational Linguistics & Cognitive Science Spring 2023

Johns Hopkins University
Teaching Assistant
Instructor: Mathias Unberath

- Machine Learning: AI System Design & Development Spring 2020

SERVICE

Organizing committees

- BlackboxNLP 2025
- The 2025 BabyLM Workshop (Co-located with EMNLP 2025)
- BlackboxNLP 2024 (Co-located with EMNLP 2024)
- The 2024 BabyLM Challenge
- The 2023 BabyLM Challenge
- The Inverse Scaling Prize (2022)

Conference chairing

- EMNLP 2025 (Publicity Chair)

Editorial responsibilities for scientific meetings

- ACL 2025 (Senior Area Chair for *Interpretability and Analysis of Models for NLP*)
- NAACL 2025 (Area Chair for *Interpretability and Analysis of Models for NLP*)
- EMNLP 2024 (Area Chair for *Interpretability and Analysis of Models for NLP*). **Outstanding Area Chair Award.**

Ad-hoc journal reviewing

- Computational Linguistics (2024)
- Journal of Memory and Language (2024)

Reviewing

- ICML (2025)
- ICLR (2025)
- NAACL (2024, 2021)
- ACL (2024, 2023, 2022, 2020)
- COLM (2024)
- NeurIPS (2024)
- EACL (2024)
- FAccT (2024)
- EMNLP (2023, 2022, 2019)
- CoNLL (2023, 2022)
- TACL (2022)
- CSCW (2021)
- COLING (2020)

SKILLS

Programming

- Languages: Python, C++, HTML, CSS, Javascript, Bash
- Machine Learning Toolkits: PyTorch (incl. HuggingFace, fairseq, sockeye), NLTK, Scikit-learn, numpy
- Version Control: DVCS (Git, Bitbucket)

Linguistic Tools

- Praat, AntConc, QGIS, Audacity

NATURAL LANGUAGES

Native: English
Proficient: French
Beginner: Hebrew