NAOMI SAPHRA

Kempner Research Fellow

Kempner Institute
Harvard University
✉ *nsaphra@nsaphra.net*
nsaphra.net

Education

2021 **Ph.D, Informatics**, *University of Edinburgh*, Edinburgh, UK.
Advisor: Adam Lopez
Thesis: *Training Dynamics of Neural Language Models.*

2015 **MSE, Computer Science**, *Johns Hopkins University*, Baltimore, MD.

2013 **B.Sc, Computer Science**, *Carnegie Mellon University*, Pittsburgh, PA.
Minor: Language Technologies

Experience

Academia

Starting 2026 **Boston University**, *Assistant Professor*, Boston, MA.
Faculty of Computing & Data Sciences (CDS).

2023–Present **Harvard University**, *Kempner Research Fellow*, Boston, MA.
Kempner Institute for the Study of Natural and Artificial Intelligence.

2021–2023 **New York University**, *Postdoctoral Researcher*, New York, NY.
Supervisor: Kyunghyun Cho.

Summer 2014 **Frederick Jelinek Memorial Workshop (JSALT)**, *Graduate Researcher*, Prague, Czech Republic.
Project: Cross-language Abstract Meaning Representation.

Summer 2012 **CLSP Summer Workshop at Johns Hopkins**, *Undergraduate Research Fellow*, Baltimore, MD.
Project: Understanding Objects in Detail with Fine-grained Attributes.

2010–2013 **Carnegie Mellon University**, *Undergraduate Research Assistant*, Pittsburgh, PA.
Supervisors: Chris Dyer and Noah Smith.

Industry

Winter 2020 **Google**, *Research Intern*, New York, NY.
Host: Dipanjan Das (Natural Language Understanding).

Summer 2017 **Koko**, *Intern*, New York, NY.
Developed classifiers for informal text at abuse-detection startup.

Summer 2015 **Google**, *Software Engineering Intern*, Mountain View, CA.
Host: Marius Pasca (Common Sense Team, Machine Intelligence).

Summer 2013 **Google**, *Software Engineering Intern*, Mountain View, CA.
Host: Eric Altendorf (Common Sense Team, Machine Intelligence).

Summer 2011 **Facebook, Inc.**, *Engineering Intern*, Palo Alto, CA.

Consulting

2022 **MosaicML**, *Consultant*, New York, NY.
Regularly advised team improving the efficiency of pretraining language models.

2022 **University of Southern California NSF grant**, *Grant Advisory Board*.
Title: *Determining Community Needs for Accessibility Tools that Facilitate Programming Education and Workforce Readiness for Persons with Disabilities*

Educational Retreats

Fall 2025 **The Aspen Meeting on Foundation Models**, *The Aspen Center for Physics*, Aspen, CO.

Spring 2017 **The Recurse Center**, New York, NY.

Summer 2012 **CLSP Workshop Summer School**, *John Hopkins University*, Baltimore, MD.

Awards & Honors

2024 **Rising Star**, *MIT Rising Stars in EECS Workshop*.

2024 **Best Reviewer**, *ICML*.

2022 **Outstanding Reviewer**, *ICLR*.

2021 **Outstanding Reviewer**, *ICLR*.

2019 **Best Poster Runner Up**, *NY Academy of Sciences - Natural Language, Dialogue & Speech*.

2019 **Daniella Sciama Award for Achievement through Adversity**, *University of Edinburgh*.
Award of £1,000.

2017 **Google Europe Scholarship for Students with Disabilities**, *Alphabet Inc*.
Award of €7,000.

2014 **Best Paper (coauthor)**, *LREC Workshop on Free/Open-Source Arabic Corpora*.

2013 **Dragon Award**, *Carnegie Mellon School of Computer Science*.
Offered by the undergraduate CS academic advisor to one graduate annually on idiosyncratic grounds.

2009 **National Merit Semifinalist**, *National Merit Scholarship Corporation*.

2009 **Nannina Rasulo Memorial Scholarship Award for Technology**, *Irvington High School*.
Award of $500.

Publications

Journals and Periodicals

2023 **Naomi Saphra**. Interpretability creationism. *The Gradient*, 2023.

2023 Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, **Naomi Saphra**, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. State-of-the-art generalisation research in NLP: a taxonomy and review. *Nature Machine Intelligence*, 2023.

2023 Michael Hu, Angelica Chen, **Naomi Saphra**, and Kyunghyun Cho. Delays, detours, and forks in the road: Latent state models of training dynamics. *Transactions of Machine Learning Research (TMLR)*, 2023.

Conference Publications

2025 Oskar van der Wal, Pietro Lesci, Max Müller-Eberstein, **Naomi Saphra**, Hailey Schoelkopf, Willem Zuidema, and Stella Biderman. Polypythias: Stability and outliers across fifty language model pre-training runs. In *International Conference on Learning Representations (ICLR)*, 2025.

2025 Divyansh Singhvi*, Diganta Misra*, Andrej Erkelens*, Raghav Jain*, Isabel Papadimitriou, and **Naomi Saphra**. Using Shapley interactions to understand how models use structure. In *Association for Computational Linguistics (ACL)*, 2025.

2025 Tian Qin, **Naomi Saphra**, and David Alvarez-Melis. Sometimes I am a tree: Data drives fragile hierarchical generalization. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

2025 USVSN Sai Prashanth*, Alvin Deng*, Kyle O'Brien*, Jyothir S V*, Mohammad Aflah Khan, Jaydeep Borkar, Christopher A. Choquette-Choo, Jacob Ray Fuehne, Stella Biderman, Tracy Ke, Katherine Lee, and **Naomi Saphra**. Recite, reconstruct, recollect: Memorization in LMs as a multifaceted phenomenon. In *International Conference on Learning Representations (ICLR)*, 2025.

2025 Michael Y. Hu, Shreyans Jain, Sangam Chaulagain, and **Naomi Saphra**. How to visualize training dynamics in neural networks. In *Blog Post Track at International Conference on Learning Representations (ICLR BlogPosts)*, 2025.

2025 Natalie Abreu, Edwin Zhang, Eran Malach, and **Naomi Saphra**. A taxonomy of transcendence. In *Conference on Language Modeling (COLM)*, 2025.

2024 Edwin Zhang, Vincent Zhu, **Naomi Saphra**, Anat Kleiman, Benjamin L. Edelman, Milind Tambe, Sham M. Kakade, and Eran Malach. Transcendence: Generative models can outperform the experts that train them. In *Neural Information Processing Systems (NeurIPS)*, 2024.

2024 **Naomi Saphra**, Eve Fleisig, Kyunghyun Cho, and Adam Lopez. First tragedy, then parse: History repeats itself in the new era of large language models. In *North American Association for Computational Linguistics (NAACL)*, 2024.

2024 Tom Sherborne, **Naomi Saphra**, Pradeep Dasigi, and Hao Peng. TRAM: Bridging Trust Regions and Sharpness Aware Minimization. In *International Conference on Learning Representations (ICLR)*, 2024. Spotlighted (top 5%).

2024 Michael Saxon, Ari Holtzman, Peter West, William Yang Wang, and **Naomi Saphra**. Benchmarks as microscopes: A call for model metrology. In *Conference on Language Modeling (COLM)*, 2024.

2024 Adir Rahamim, **Naomi Saphra**, Sara Kangaslahti, and Yonatan Belinkov. Fast forwarding low-rank training. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

2024 Victoria R. Li*, Yida Chen*, and **Naomi Saphra**. ChatGPT doesn't trust Chargers fans: Guardrail sensitivity in context. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

2024 Angelica Chen, Ravid Schwartz-Ziv, Kyunghyun Cho, Matthew Leavitt, and **Naomi Saphra**. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. In *International Conference on Learning Representations (ICLR)*, 2024. Spotlighted (top 5%).

2024 Ian Berlot-Attwell, Kumar Krishna Agrawal, A. Michael Carrell, Yash Sharma, and **Naomi Saphra**. Attribute diversity determines the systematicity gap in VQA. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

2024 Zachary Ankner, **Naomi Saphra**, Davis Blalock, Jonathan Frankle, and Matthew L. Leavitt. Dynamic masking rate schedules for MLM pretraining. In *European Association for Computational Linguistics (EACL)*, 2024. Accepted as oral presentation.

2023 Jeevesh Juneja, Rachit Bansal, Kyunghyun Cho, João Sedoc, and **Naomi Saphra**. Linear Connectivity Reveals Generalization Strategies. In *International Conference on Learning Representations (ICLR)*, 2023.

2022 Josef Valvoda, **Naomi Saphra**, Jonathan Rawski, Ryan Cotterell, and Adina Williams. Learning Transductions to Test Systematic Compositionality. In *International Conference on Computational Linguistics (COLING)*, 2022.

2022 Thibault Sellam, Steve Yadlowsky, Ian Tenney, Jason Wei, **Naomi Saphra**, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Raluca Turc, Jacob Eisenstein, Dipanjan Das, and Ellie Pavlick. The MultiBERTs: BERT Reproductions for Robustness Analysis. In *International Conference on Learning Representations (ICLR)*, 2022. Spotlighted (top 5%).

2021 Jennifer C. White, Tiago Pimentel, **Naomi Saphra**, Adina Williams, and Ryan Cotterell. A Non-Linear Structural Probe. In *North American Association for Computational Linguistics (NAACL)*, 2021.

2020 **Naomi Saphra** and Adam Lopez. LSTMs Compose—and Learn—Bottom-Up. In *Findings of Empirical Methods in Natural Language Processing (EMNLP Findings)*, 2020.

2020 Mohammad Tahaei, Kami Vaniea, and **Naomi Saphra**. Understanding privacy-related questions on stack overflow. In *Conference on Human Factors in Computing Systems (CHI)*, 2020.

2020 Tiago Pimentel*, **Naomi Saphra***, Adina Williams, and Ryan Cotterell. Pareto Probing: Trading Off Accuracy for Complexity. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

2019 **Naomi Saphra** and Adam Lopez. Understanding Learning Dynamics Of Language Models with SVCCA. In *North American Association for Computational Linguistics (NAACL)*, 2019.

2015 **Naomi Saphra** and Adam Lopez. AMRICA: an AMR Inspector for Cross-language Alignments. In *North American Association for Computational Linguistics (NAACL) (demos)*, 2015.

2014 Andrea Vedaldi, Siddharth Mahendran, Stavros Tsogkas, Subhransu Maji, Ross Girshick, Juho Kannala, Esa Rahtu, Iasonas Kokkinos, Matthew B. Blaschko, David Weiss, Ben Taskar, Karen Simonyan, **Naomi Saphra**, and Sammy Mohamed. Understanding Objects in Detail with Fine-grained Attributes. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.

Workshop Publications

2024 **Naomi Saphra*** and Sarah Wiegreffe*. Mechanistic? In *EMNLP BlackboxNLP Workshop*, 2024. Oral presentation (Top 10%).

2023 Yash Gondhalekar, Sultan Hassan, **Naomi Saphra**, and Sambatra Andrianomena. Towards out-of-distribution generalization in large-scale astronomical surveys: robust networks learn similar representations. In *NeurIPS workshop on Machine Learning and the Physical Sciences*, 2023.

2022 Bingchen Zhao*, Yuling Gu*, Jessica Zosa Forde, and **Naomi Saphra**. One Venue, Two Conferences: The Separation of Chinese and American Citation Networks. In *NeurIPS Workshop on Cultures of AI and AI for Culture*, 2022.

2019 **Naomi Saphra** and Adam Lopez. Sparsity emerges naturally in neural language models. In *ICML Workshop on Identifying and Understanding Deep Learning Phenomena*, 2019.

2019 Kate McCurdy and **Naomi Saphra**. Carbon AI and the concentration of computational work. In *Challenging the Work Society: an interdisciplinary summit*, 2019.

2016 **Naomi Saphra** and Adam Lopez. Evaluating Informal-Domain Word Representations with UrbanDictionary. In *ACL Workshop on Evaluating Vector Space Representations for NLP (RepEval)*, 2016.

2014 Nathan Schneider, Brendan O'Connor, **Naomi Saphra**, David Bamman, Manaal Faruqui, Noah A. Smith, Chris Dyer, and Jason Baldridge. A framework for (under) specifying dependency syntax without overloading annotators. In *ACL Linguistic Annotation Workshop*, 2014.

2014 Ryan Cotterell, Adithya Renduchintala, **Naomi Saphra**, and Chris Callison-Burch. An Algerian Arabic-French Code-Switched Corpus. In *LREC Workshop on Free/Open-Source Arabic Corpora*, 2014. Best paper award.

Preprints and Preliminary Manuscripts

2025 Kaden Zheng, Sonja Johnson-Yu, Satpreet Harcharan Singh, Denis Turcu, Federico Pedraja, Pratyusha Sharma, Naomi Saphra, Nathaniel Sawtell, and Kanaka Rajan. Keypoint annotation for electrocommunication source separation with PIKAChu and RAIChu. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems workshop: AI for non-human animal communication*, 2025.

2025 Rosie Zhao*, Tian Qin*, David Alvarez-Melis, Sham Kakade, and **Naomi Saphra**. Distributional scaling of emergent capabilities, 2025. Presented at 2024 NeurIPS Workshop on Scientific Methods for Understanding Deep Learning.

2025 Satpreet Harcharan Singh, Sonja Johnson-Yu, Zhouyang Lu, Aaron Walsman, Federico Pedraja, Denis Turcu, Pratyusha Sharma, Naomi Saphra, Nathaniel Sawtell, and Kanaka Rajan. Understanding electro-communication and electro-sensing in weakly electric fish using multi-agent deep reinforcement learning. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems workshop: AI for non-human animal communication*, 2025.

2025 Satpreet Harcharan Singh, Sonja Johnson-Yu, Zhouyang Lu, Aaron Walsman, Federico Pedraja, Denis Turcu, Pratyusha Sharma, Naomi Saphra, Nathaniel Sawtell, and Kanaka Rajan. Proposal: Deciphering electrocommunication with MARL and unsupervised machine translation. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems workshop: AI for non-human animal communication*, 2025.

2025 Victoria R. Li*, Jenny Kaufmann*, Martin Wattenberg, David Alvarez-Melis, and **Naomi Saphra**. Can interpretation predict behavior on unseen data? In *arXiv*, 2025. Presented at *2024 NeurIPS Workshop on Scientific Methods for Understanding Deep Learning* and *2025 NeurIPS Workshop on Mechanistic Interpretability*.

2025 Millicent Li, Alberto Mario Ceballos Arroyo, Giordano Rogers, Naomi Saphra, and Byron C. Wallace. Do natural language descriptions of model activations convey privileged information?, 2025. Presented at *2025 NeurIPS Workshop on Mechanistic Interpretability*.

2025 Sara Kangaslahti, Elan Rosenfeld, and **Naomi Saphra**. Loss in the crowd: Hidden breakthroughs in language model training. In *arXiv*, 2025. Spotlighted at *2024 ICML Workshop on Mechanistic Interpretability*.

2024 Sonja Johnson-Yu, Satpreet Harcharan Singh, Federico Pedraja, Denis Turcu, Pratyusha Sharma, **Naomi Saphra**, Nathaniel Sawtell, and Kanaka Rajan. Understanding biological active sensing behaviors by interpreting learned artificial agent policies. In *Workshop on Interpretable Policies in Reinforcement Learning @RLC-2024*, 2024. Accepted as oral presentation.

2017 Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, **Naomi Saphra**, Swabha Swayamdipta, and Pengcheng Yin. DyNet: The dynamic neural network toolkit, 2017.

Media Coverage

Sep 2025 **Quanta Magazine**.
To Understand AI, Watch How It Evolves

Aug 2025 **The Register**.
ChatGPT hates LA Chargers fans

April 2024 **Harvard Gazette**.
Why AI fairness conversations must include disabled people

May 2023 **Gary Marcus's Humans vs. Machines Podcast**.
S4E4: Can AI Make You Laugh?

July 2017 **New Scientist**.
Donate your voice so Siri doesn't just work for white men

Talks

Invited event talks

Dec 2025 **QCon.AI**, New York, NY.

Nov 2025 **Princeton Center for Theoretical Science Workshop: The Physics of John Hopfield**, Princeton, NJ.

Oct 2025 **The Aspen Meeting on Foundation Models**, Aspen, CO.

July 2025 **ICML 2025 Workshop on Assessing World Models (*Keynote*)**, Vancouver, Canada.
And Nothing Between: Using Categorical Differences to Understand and Predict Model Behavior

July 2025 **Prague Workshop on Neural Networks and Reasoning**, Prague, Czechia (Remote).

June 2025 **Computational Linguistics and Linguistic Theory (COLT) Symposium on Emergence**, Barcelona, Spain.

May 2025 **International Conference on the Mathematics of Neuroscience and AI (NeuroMonster)**, Split, Croatia.

Mar 2025 **Safety Workshop – Principles of Intelligent Behavior in Biological and Social Systems**, Boston, MA.

Feb 2025 **Simons Institute Workshop: LLMs, Cognitive Science, Linguistics, and Neuroscience**, Berkeley, CA.

Sept 2024 **Simons Institute Workshop: Transformers as a Computational Model**, Berkeley, CA.

June 2022 **3rd Neural Scaling Laws Workshop (*Keynote*)**, Manoir Saint-Sauveur, Quebec.
Sources of Variance in Pretraining and Finetuning

Nov 2020 **EMNLP QueerInAI Social**, Seattle, WA (Remote).

June 2020 **Pydatafest Amsterdam (*Keynote*)**, Amsterdam, Netherlands (Remote).
Accessible Means Hackable

Jan 2019 **Understanding & Analyzing Neural Networks Workshop**, Amsterdam, Netherlands.

Panels

Dec 2025 **NeurIPS Tutorials**, San Diego, CA.
Tutorial on Benchmarking Practices

Aug 2025 **New England Mechanistic Interpretability Workshop (NEMI)**, Boston, MA.
Bridging the Gap: From Lowest-Level Mechanisms to High-Level Behaviors

July 2025 **ICML Actionable Interpretability Workshop**, Vancouver, Canada.
Actionable Interpretability

Dec 2024 **Aethos Global Summit on Open Problems for AI**, Boston, MA.
Ethics & Society (Micro-ethics)

Oct 2024 **COLM MLR@Penn Workshop on Foundation Models**, Philadelphia, PA.
Emerging Trends in Open Foundation Model Development

Sept 2024 **Simons Institute Workshop: Transformers as a Computational Model**, Berkeley, CA.
What can theory offer to the design and use of LLMs?

July 2024 **ICML Mechanistic Interpretability Workshop**, Vienna, Austria.
Interpretability Panel Discussion

July 2024 **ICML Queer and {Dis}ability in AI Social**, Vienna, Austra.
Human-AI Interactions and Underrepresented Communities

May 2024 **ICLR Interpretability Social**, Vienna, Austria.
Interpretability Panel Discussion

June 2020 **NAACL D&I Sessions**, Remote.
D&I Session: Inclusivity in Conferences

Aug 2019 **ACL Blackbox NLP Workshop**, Florence, Italy.
Blackbox NLP Panel Discussion

Other invited talks

Oct 2025 **Carnegie Mellon University**, Pittsburgh, PA.
LTI Colloquium

Sept 2025 **New York University**, New York, NY.
NYU Text-As-Data Seminar Series

July 2025 **ML Collective**, Remote.
Deep Learning Concepts and Trends

April 2025 **Institute for Artificial Intelligence and Fundamental Interactions (IAIFI)**, Cambridge, MA.
Meeting on interpretability for science

Apr 2025 **UC Santa Barbara**, Santa Barbara, CA (Remote).
NLP Lab Seminar

Nov 2024 **Spotify, Inc.**, Boston, MA.

May 2024 **Stanford University**, Palo Alto, CA (Remote).
Stanford NLP Seminar

March 2024 **Massachusetts Institute of Technology**, Cambridge, MA.
MIT Embodied Intelligence Seminar Series

Feb 2024 **University of Massachusetts – Amherst**, Amherst, MA.
UMass NLP Seminar Series

Nov 2023 **Carnegie Mellon University**, Pittsburgh, PA.
Machine Learning Faculty / Duolingo Seminar Series

Aug 2023 **Microsoft Research**, Montreal, Canada (Remote) and New York, NY.
MSR Montreal Seminar Series

June 2023 **Heriot-Watt University**, Edinburgh, UK (Remote).
Lab for AI Verification Speaker Series

March 2023 **University of Edinburgh**, Edinburgh, UK.
NLP Seminar Series

March 2023 **University of Copenhagen**, Copenhagen, Denmark.
University of Copenhagen NLP Seminar Series

Feb 2023 **Georgetown University**, Washington, DC.
Nathan Schneider lab

Jan 2023 **Massachusetts Institute of Technology**, Boston, MA (Remote).
The Center for Biological & Computational Learning speaker series

July 2022 **Oracle**, Boston, MA (Remote).
Machine Learning Seminar

June 2022 **Stanford University**, Palo Alto, CA (Remote).
Stanford NLP Seminar

June 2022 **UC Irvine**, Irvine, CA.
Sameer Singh lab

June 2022 **University of Southern California - Information Sciences Institute**, Irvine, CA.
USC ISI Natural Language Seminar

Feb 2022 **University College London**, London, UK.

Nov 2020 **UC Berkeley**, Berkeley, CA (Remote).
Berkeley NLP Seminar

May 2020 **Brown University**, Providence, RI (Remote).
Brown NLP Seminar

Sept 2019 **Element AI**, London, UK.

Aug 2019 **Allen Institute for AI**, Seattle, WA.

May 2019 **City University of New York**, New York, NY.
Kyle Gorman lab

Outreach

Dec 2025 **Westchester Public Libraries speaker series**, Tuckahoe, NY (Remote).
Rules for Understanding Language Models

Sept 2025 **The Forum at Newport**, Newport, RI.
Naval War College event for junior officers

July 2023 **HackNY Fellows Speaker Series**, New York, NY.

Jan 2023 **Westchester Public Libraries speaker series**, Tuckahoe, NY (Remote).
Hacking Disability: Accessible and Adaptable Tech

March 2020 **!!Con West**, Santa Cruz, CA.
Get Hooked on Pytorch Hooks!

Sept 2019 **Challenging the Work Society**, London, UK.
Carbon AI and the Concentration of Computational Work (jointly presented with Kate McCurdy)

Feb 2019 **Bright Club (The Stand – Edinburgh Comedy Club)**, Edinburgh, UK.
Paying the Panopticon (standup comedy)

Teaching

Teaching Support

2021, 2023 **Center for Data Science Capstone**, *New York University*.
Project Mentor

2019 **Probabilistic Modeling & Reasoning**, *University of Edinburgh*.
Tutor (Teaching recitations)

2017–2019 **Machine Learning & Pattern Recognition**, *University of Edinburgh*.
Tutor (Teaching recitations)

2016 **Informatics Research Review**, *University of Edinburgh*.
Tutor (Teaching recitations)

2013 **Natural Language Processing**, *Johns Hopkins University*.
Course Assistant (Grading)

2008–2009 **Ancient Greek**, *Irvington High School*.
Teacher's Assistant (Grading)

Guest Lectures

Jan 2026 **Stanford University**, Palo Alto, CA.
Graduate seminar on philosophical and conceptual issues in AI

Apr 2025 **Brown University**, Providence, RI.
Interpretability course guest lecture – Understanding Training Dynamics

Nov 2024 **Korea Advanced Institute of Science & Tech (KAIST)**, Daejeon, South Korea (Remote).
ML for NLP (CS475) guest lecture – What determines LM training outcomes?

Jan 2022 **NYU AI School**, New York, NY (Remote).
Mathematical Fundamentals of AI

Master's Thesis Supervision

2021 **University of Amsterdam MSc Thesis**, Sylke Gosen.
Understanding Language Models through Perturbed Datasets.
Co-advised with Dieuwke Hupkes and Jaap Jumelet.

2018 **University of Edinburgh MSc Thesis**, Alp Ozkan.
Combined Application of Pruning and Growing Approaches for Neural Networks.
Co-advised with Adam Lopez.

2018 **University of Edinburgh MSc Thesis**, Yekun Chai.
Discovering Spelling Variants on Urban Dictionary.
Co-advised with Adam Lopez.

Undergraduate Supervision

2025 **Harvard Kempner KURE Fellowship**, Simon Ma.

2025 **Harvard Kempner KRANIUM Fellowship**, Laasya Nagumalli.

2025 **Harvard HCRP Fellowship**, Kaden Zheng.

2024 **Harvard Kempner KURE Fellowship**, Victoria Li.
Co-advised with David Alvarez-Melis.

2024 **Harvard PRISE Fellowship**, Victoria Li.
Co-advised with David Alvarez-Melis.

2024 **Harvard Kempner KRANIUM Fellowship**, Anne Mykland.
Co-advised with David Alvarez-Melis.

2024 **Harvard Kempner KURE Fellowship**, Helen He.
Co-advised with David Alvarez-Melis.

Service

Organization

2024 **Organizing committee**, *ICML Workshop on High-dimensional Learning Dynamics (HiLD): The Emergence of Structure and Reasoning*, Vienna, Austria.

2023 **Organizing committee**, *ACL Workshop on Representation Learning for NLP (RepL4NLP)*, Toronto, Canada.

2022 **Organizing committee**, *EMNLP Workshop on Analyzing and interpreting neural networks for NLP (BlackboxNLP)*, Abu Dhabi, United Arab Emirates.

2021 **Organizing committee**, *ACL Workshop on Representation Learning for NLP (RepL4NLP)*, Remote.

2014 **Social co-chair**, *Association for Computational Linguistics (ACL)*, Baltimore, MD.

Session moderation

Aug 2025 **Roundtable lead**, *New England Mechanistic Interpretability Workshop (NEMI)*, Boston, MA.
Training Dynamics Roundtable.

Dec 2023 **Panel Moderator**, *NeurIPS Negative Results Workshop*, New Orleans, LA.
Negative Results Panel Discussion.

June 2022 **Session chair**, *Association for Computational Linguistics (ACL)*, Dublin, Ireland.
Interpretability & Analysis Track.

Outreach & Inclusion

2023 **Project Mentor**, *ACL Student Research Workshop*, Toronto, Canada.

2020 **Accessibility Subcommittee**, *Association for Computational Linguistics (ACL)*, Remote.

2019-2020 **Disability representative**, *University of Edinburgh Staff Pride Network*, Edinburgh, UK.

2012–2014 **NACLO volunteer**, *Johns Hopkins University and Carnegie Mellon University*.
Puzzle tester and local organizer for the North American Computational Linguistics Olympiad.

Refereeing

Senior Area Chair

2025 **North American Association for Computational Linguistics (NAACL)**.

Area Chair / Action Editor

2025–Present **International Conference on Learning Representations (ICLR)**.

2024–Present **Association for Computational Linguistics (ACL) Rolling Review**.
Includes service for EMNLP, ACL, NAACL, AACL, and EACL.

2024–Present **Conference on Language Models (COLM)**.

2025 **NeurIPS Mechanistic Interpretability Workshop**.

2024 **NeurIPS Workshop on Interpretable AI: Past, Present and Future**.

2021–2024 **Empirical Methods in Natural Language Processing (EMNLP)**.

2024 **Language Resources and Evaluation Conference (LREC)**.

2022 **Asian Association for Computational Linguistics (AACL)**.

Journal Reviewing

Computational Linguistics (CL).

Journal of Machine Learning Research (JMLR).

Conference Reviewing

2025–Present **Neural Information Processing Systems (NeurIPS) – Position Paper Track**.

2025–Present **International Conference on Machine Learning (ICML) – Position Paper Track**.

2020–2024 **International Conference on Machine Learning (ICML)**.
2024 Best Reviewer

2021–2024 **Neural Information Processing Systems (NeurIPS)**.

2021–2024 **International Conference on Learning Representations (ICLR)**.
2022 Outstanding Reviewer
2021 Outstanding Reviewer

2024 **COGSCI**.

2021–2023 **Association for Computational Linguistics (ACL) Rolling Review**.

2021–2023 **European Association for Computational Linguistics (EACL)**.

2019–2023 **Association for Computational Linguistics (ACL)**.

2019–2021 **North American Association for Computational Linguistics (NAACL)**.

2017–2020 **Empirical Methods in Natural Language Processing (EMNLP)**.

Workshop and Competition Reviewing

2025 **Cognitive Interpretability Workshop**, *NeurIPS*.

2024 **Mechanistic Interpretability Workshop**, *ICML*.

2018–2023 **Widening NLP (WiNLP)**, *ACL*.

2019–2023 **Analyzing and interpreting neural networks for NLP (BlackboxNLP)**, *ACL*.

2021–2023 **Negative Results Workshop**, *NeurIPS*.

2022 **Inverse Scaling Prize**.

2020 **The SIGNLL Conference on Computational Natural Language Learning (CoNLL)**, *ACL*.

2017–2019 **Learning With Limited Data Workshop (LLD)**, *ICLR*.

2017 **Women in Machine Learning**, *NeurIPS*.

2017 **Representation Evaluation for NLP (RepL4NLP)**, *ACL*.