NAOMI SAPHRA

Curriculum Vitae

Kempner Institute
Harvard University
✉ nsaphra@nsaphra.net
⌂ nsaphra.net
○ nsaphra   🐦 nsaphra

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.
Mentors: Adam Lopez, Sanjeev Khudanpur, Raman Arora
Obtained as a PhD student before transferring to Edinburgh.

2013   **B.Sc, Computer Science**, *Carnegie Mellon University*, Pittsburgh, PA.
Mentors: Noah Smith and Chris Dyer
Minor: Language Technologies

Spring 2017   **Sabbatical**, *The Recurse Center*, New York, NY.
Experimental unstructured educational programming retreat.

Work

2023–Present   **Kempner Institute at Harvard University**, *Kempner Research Fellow*, Boston, MA.

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

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

Winter 2020   **Google**, *Research Intern*, New York, NY.
Mentors: Dipanjan Das, Ian Tenney, Jasmijn Bastings, Thibault Sellam.
Topic: Variance in pretraining dynamics across random seeds.

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

Summer 2015   **Google**, *Software Engineering Intern*, Mountain View, CA.
Mentor: Marius Pasca (Common Sense Team, Machine Intelligence).
Topic: Unsupervised discovery of paraphrases.

Summer 2012   **Frederick Jelinek Memorial Workshop (JSALT)**, *Graduate Researcher*, Prague, Czech Republic.
Project: Cross-language Abstract Meaning Representation.

Summer 2013   **Google**, *Software Engineering Intern*, Mountain View, CA.
Mentor: Eric Altendorf (Common Sense Team, Machine Intelligence).
Topic: Integrating models of real-world knowledge into entity linking systems.

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

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

Summer 2011   **Facebook, Inc.**, *Engineering Intern*, Palo Alto, CA.
Mentor: Jen Burge (Engagement Team).
Topic: Improving the People You May Know model.

Publications

Conference Publications

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. Acceptance rate: 32%.

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. Acceptance rate: 33%.

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. Acceptance rate: 32%.

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. Acceptance rate: 26%.

2020   **Naomi Saphra** and Adam Lopez. LSTMs Compose—and Learn—Bottom-Up. In *Findings of Empirical Methods in Natural Language Processing (EMNLP)*, 2020. Acceptance rate: 38% for EMNLP and Findings combined.

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. Acceptance rate: 29%.

2020   Tiago Pimentel*, **Naomi Saphra*** (equal contribution), Adina Williams, and Ryan Cotterell. Pareto Probing: Trading Off Accuracy for Complexity. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2020. Acceptance rate: 22%.

2019   **Naomi Saphra** and Adam Lopez. Understanding Learning Dynamics Of Language Models with SVCCA. In *North American Association for Computational Linguistics (NAACL)*, 2019. Acceptance rate: 23%.

2015   **Naomi Saphra** and Adam Lopez. AMRICA: an AMR Inspector for Cross-language Alignments. In *North American Association for Computational Linguistics (NAACL) (demos)*, 2015. Acceptance rate: 26%.

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. Acceptance rate: 30%.

Journal Publications

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.

Workshop Publications

2023 Divyansh Singhvi*, Andrej Erkelens*, Raghav Jain*, Diganta Misra, and **Naomi Saphra**. Shapley interactions for complex feature attribution. In *NeurIPS Workshop on Attributing Model Behavior at Scale (ATTRIB)*, 2023.

2023 Yash Gondhalekar, Sultan Hassan, **Naomi Saphra**, and Sambatra Andrianomena. Towards out-of-distribution generalization: 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*, October 2022.

2019 **Naomi Saphra** and Adam Lopez. Sparsity emerges naturally in neural language models. In *ICML Workshop on Identifying and Understanding Deep Learning Phenomena (Deep Phenomena)*, 2019. Acceptance rate: 49%.

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. Acceptance rate: 67%.

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. Acceptance rate: 50%.

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. Acceptance rate: 53%. Received best paper award.

Preprints and Works Under Submission

2023 **Naomi Saphra**, Eve Fleisig, Kyunghyun Cho, and Adam Lopez. First tragedy, then parse: History repeats itself in the new era of large language models. *arXiv preprint*, 2023.

2023 Tom Sherborne, **Naomi Saphra**, Pradeep Dasigi, and Hao Peng. TRAM: Bridging trust regions and sharpness aware minimization. *arXiv preprint*, 2023.

2023 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. *arXiv preprint*, 2023.

2023 Ian Berlot-Attwell, A. Michael Carrell, Kumar Krishna Agrawal, Yash Sharma, and **Naomi Saphra**. Attribute diversity determines the systematicity gap in VQA. *arXiv preprint*, 2023.

2023 Zachary Ankner, **Naomi Saphra**, Davis Blalock, Jonathan Frankle, and Matthew L. Leavitt. Dynamic masking rate schedules for MLM pretraining. *arXiv preprint*, 2023.

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. *arXiv preprint*, 2017.

Awards & Grants

2022 **Advisory Board for University of Southern California grant**, *National Science Foundation*.
Title: *Convergence Accelerator Track H: Determining Community Needs for Accessibility Tools that Facilitate Programming Education and Workforce Readiness for Persons with Disabilities*
Award of $698,161.

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

2021	**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 academic advisor to one CS 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.

Teaching

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)

Supervision

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

2018 **University of Edinburgh MSc.**, Alp Ozkan.
Combined Application of Pruning and Growing Approaches for Neural Networks.

2018 **University of Edinburgh MSc.**, Yekun Chai.
Discovering Spelling Variants on Urban Dictionary.

Talks

Invited plenary events

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

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

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

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

Jan 2019	**Understanding & Analyzing Neural Networks Workshop**, Amsterdam, Netherlands.
	Understanding Learning Dynamics Of Language Models with SVCCA

Nov 2023	**Carnegie Mellon University**, Pittsburgh, PA.
	Machine Learning Faculty 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	**Harvard University**, Boston, MA (Remote).
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	**Toyota Technical Institute of Chicago**, Chicago, IL.
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
Oct 2022	**Formal Languages and Neural Networks (FLaNN) Discord**, Remote.
July 2022	**Oracle**, Boston, MA (Remote).
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
March 2022	**Formal Languages and Neural Networks (FLaNN) Discord**, Remote.
Feb 2022	**University College London**, London, UK.
Nov 2020	**UC Berkeley**, Berkeley, CA (Remote).
	Berkeley NLP Seminar
Oct 2020	**NLP With Friends**, Remote.
June 2020	**New York University**, New York, NY (Remote).
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

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

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	**The Stand Edinburgh Comedy Club**, Edinburgh, UK.
	Paying the Panopticon (standup comedy)
Oct 2012	**Carnegie Mellon TechNights**, Pittsburgh, PA.
	ML and NLP Guest Lectures (program for middle school girls)

Service

General

2023	**Organizing committee**, *8th Workshop on Representation Learning for NLP (RepL4NLP) at ACL*, Toronto, Canada.
2022	**Organizing committee**, *Analyzing and interpreting neural networks for NLP (BlackboxNLP) at EMNLP*, Abu Dhabi, United Arab Emirates.
2022	**Session Chair**, *Association for Compuational Linguistics (ACL)*, Dublin, Ireland.
2021	**Organizing committee**, *6th Workshop on Representation Learning for NLP (RepL4NLP) at ACL*, Remote.
2014	**Social co-chair**, *Association for Compuational Linguistics (ACL)*, Baltimore, MD.
2012–2014	**NACLO volunteer**, *Johns Hopkins University and Carnegie Mellon University*.
	Tested puzzles for the national North American Computational Linguistics Olympiad.
	Organized local high school competition.

Diversity, Equity, & Inclusion

2023	**Project Mentor**, *ACL Student Research Workshop*, Toronto, Canada.
2020	**Panelist - D&I Session: Inclusivity in Conferences**, *North American Association for Compuational Linguistics (NAACL)*, Remote.
2020	**Accessibility Subcomittee**, *Association for Compuational Linguistics (ACL)*, Remote.
2019-2020	**Disability representative**, *University of Edinburgh Staff Pride Network*, Edinburgh, UK.

Refereeing

Area Chair

2021–Present	**Empirical Methods in Natural Language Processing (EMNLP)**, 2021, 2022, 2023.
2024	**European Association for Computational Linguistics (EACL)**, 2024.
2024	**Language Resources and Evaluation Conference (LREC)**, 2024.
2022	**Asian Association for Computational Linguistics (AACL)**, 2022.

Journal Reviewing

2023–Present	**Journal of Machine Learning Research (JMLR)**.
2022–Present	**Computational Linguistics**.

Conference Reviewing

2021–Present	**International Conference on Learning Representations (ICLR)**, 2021, 2022, 2023, 2024.
	2022 Outstanding Reviewer
	2021 Outstanding Reviewer
2017–Present	**Neural Information Processing Systems (NeurIPS)**, 2017 (secondary), 2021, 2022, 2023.
2020–Present	**International Conference on Machine Learning (ICML)**, 2020, 2023.
2021–2023	**Association for Computational Linguistics (ACL) Rolling Review**.

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

2019–2023 **Association for Computational Linguistics (ACL)**, 2019, 2020, 2021, 2023.

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

2017–2020 **Empirical Methods in Natural Language Processing (EMNLP)**, 2017, 2018, 2019, 2020.

Workshop and Competition Reviewing

2018–Present **Widening NLP (WiNLP)**, *ACL*, 2018, 2019, 2020, 2023.

2019–Present **Analyzing and interpreting neural networks for NLP (BlackboxNLP)**, *ACL*, 2019, 2021, 2023.

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

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, 2019.

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

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