

John Hewitt

Assistant Professor Department of Computer Science, Columbia University

Visiting Researcher Google DeepMind

jh5020@columbia.edu <https://cs.columbia.edu/~johnhew/>

EDUCATION

Stanford University 2018–2024

Ph.D. Computer Science.

University of Pennsylvania 2014–2018

B.S.E. Computer and Information Science.

PUBLICATIONS

Neologism Learning for Controllability and Self-Verbalization.

John Hewitt, Oyvind Tafjord, Robert Geirhos, Been Kim.

Preprint. 2025.

Because we have LLMs, we Can and Should Pursue Agentic Interpretability

Been Kim, **John Hewitt**, Neel Nanda, Noah Fiedel, Oyvind Tafjord.

Preprint. 2025.

We Can’t Understand AI Using our Existing Vocabulary.

John Hewitt, Robert Geirhos, Been Kim.

International Conference on Machine Learning (Position Paper Track.) 2025.

Instruction Following without Instruction Tuning.

John Hewitt, Nelson F. Liu, Christopher D. Manning, Percy Liang.

Preprint 2025.

Closing the Curious Case of Neural Text Degeneration

Matthew Finlayson, **John Hewitt**, Alexander Koller, Swabha Swayamdipta, Ashish Sabharwal.

In *International Conference on Learning Representations*. Vienna. May 2024.

Model Editing with Canonical Examples

John Hewitt, Sarah Chen, Lanruo Lora Xie, Edward Adams, Percy Liang, Christopher D. Manning.

Preprint 2024.

A non-archival version received Runner-Up Best Paper at the R0-FoMo Workshop @ NeurIPS 2023

Backpack Language Models

John Hewitt, John Thickstun, Christopher D. Manning, and Percy Liang.

In *Proceedings of the Conference of the Association for Computational Linguistics*. Toronto, Canada. July 2023.

Outstanding Paper Award.

Lost in the Middle: How Language Models Use Long Contexts

Nelson F. Liu, Kevin Lin, **John Hewitt**, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, Percy Liang.

Transactions of the Association for Computational Linguistics 2023.

Chinese Character-Level Backpack Language Models

Hao Sun, **John Hewitt**.

In *BlackBoxNLP: Analyzing and Interpreting Neural Networks for NLP Workshop*. Singapore. December, 2023.

Truncation Sampling as Language Model Desmoothing

John Hewitt, Christopher D. Manning, and Percy Liang.

In *Findings of the Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, UAE. November 2022.

JamPatoisNLI: A Jamaican Patois Natural Language Inference Dataset

Ruth-Ann Hazel Armstrong, **John Hewitt**, and Christopher D. Manning.

In *Findings of the Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, UAE. November 2022.

Conditional probing: measuring usable information beyond a baseline

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher D. Manning.

In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Punta Cana, Dominican Republic.

November 2021.

On the Opportunities and Risks of Foundation Models

Bommasani et. al. **John Hewitt**: co-lead, Interpretability section.

In *ArXiv*. Virtual. August 2021.

Refining Targeted Syntactic Evaluation of Language Models

Benjamin Newman, Kai-Siang Ang, Julia Gong, and **John Hewitt**.

In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. Virtual. June 2021.

Probing artificial neural networks: Insights from neuroscience

Anna Ivanova, **John Hewitt**, and Noga Zaslavsky.

In *Proceedings of the Brain2AI Workshop*. Virtual. May 2021.

RNNs can generate bounded hierarchical languages with optimal memory

John Hewitt, Michael Hahn, Surya Ganguli, Percy Liang, and Christopher D. Manning.

In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Virtual. November 2020.

The EOS Decision and Length Extrapolation

Benjamin Newman, **John Hewitt**, Percy Liang, and Christopher D. Manning.

In *BlackBoxNLP: Analyzing and Interpreting Neural Networks for NLP Workshop*. Virtual. November 2020

Outstanding Paper Award.

Emergent Linguistic Structure in Artificial Neural Networks Trained by Self-Supervision

Christopher D. Manning, Kevin Clark, **John Hewitt**, Urvashi Khandelwal, and Omer Levy.

Proceedings of the National Academy of Sciences. June 2020.

Finding Universal Grammatical Relations in Multilingual BERT

Ethan Chi, **John Hewitt**, and Christopher D. Manning.

In *Proceedings of the Conference of the Association for Computational Linguistics..* Virtual. July 2020.

Designing and Interpreting Probes with Control Tasks

John Hewitt and Percy Liang.

In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Hong Kong, China. November 2019.

Runner Up Best Paper Award.

A Structural Probe for Finding Syntax in Word Representations

John Hewitt and Christopher D. Manning.

In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis. June 2019.

Simple, Fast, Accurate Intent Classification and Slot Labeling for Goal-Oriented Dialogue Systems

Arshit Gupta* and **John Hewitt*** and Katrin Kirchhoff.

In *Proceedings of the SIGDIAL 2019 Conference*. Stockholm, Sweden. September 2019.

*: Equal contribution; authors listed alphabetically.

A Distributional and Orthographic Aggregation Model for English Derivational Morphology

Daniel Deutsch*, **John Hewitt*** and Dan Roth.

In *Proceedings of the Conference of the Association for Computational Linguistics*. Melbourne, Australia. July 2018.

*: Equal contribution; authors listed alphabetically.

Learning Translations via Images with a Massively Multilingual Image Dataset

John Hewitt*, Daphne Ippolito*, Brendan Callahan, Reno Kriz, Derry Tanti Wijaya and Chris Callison-Burch.

In *Proceedings of the Conference of the Association for Computational Linguistics*. Melbourne, Australia. July 2018.

*: Equal contribution; authors listed alphabetically.

XNMT: The eXtensible Neural Machine Translation Toolkit

Graham Neubig, Matthias Sperber, Xinyi Wang, Matthieu Felix, Austin Matthews, Sarguna Padmanabhan, Ye Qi, Devendra Singh Sachan, Philip Arthur, Pierre Godard, **John Hewitt**, Rachid Riad, and Liming Wang.

In *Conference of the Association for Machine Translation in the Americas (AMTA) Open Source Software Showcase*. Boston. March 2018.

Learning Translations via Matrix Completion

Derry Tanti Wijaya, Brendan Callahan, **John Hewitt**, Xiao Ling, Marianna Apidianaki, and Chris Callison-Burch.

In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark. September 2017.

Automatic Construction of Morphologically-Motivated Translation Models for Highly Inflected Low-Resource Languages

John Hewitt, Matt Post, David Yarowsky.

In *Proceedings of the Conference of the Association for Machine Translation in the Americas*. Austin. October 2016.

RESEARCH EXPERIENCE

- Google DeepMind August 2024–
Visiting Researcher, *with Been Kim*
- Stanford University September 2018–2024
PhD Researcher, *with Chris Manning and Percy Liang*
- DeepMind June 2022–October 2022
Research Scientist Intern, *with Aida Nematzadeh and Adhiguna Kuncoro*
- Google AI September 2020–February 2021
Research Intern, *with Vincent Zhao and Kelvin Guu*
- Amazon AI May 2018–September 2018
Applied Scientist Intern, *with Katrin Kirchhoff and Arshit Gupta*
- University of Pennsylvania 2016–2018
Research Assistant, *with Chris Callison-Burch*
- Johns Hopkins University May 2015–May 2016
Research Assistant, *with David Yarowsky and Matt Post*

TEACHING & MENTORING EXPERIENCE

- *Instructor*. Columbia. COMS 4705: Natural Language Processing. 2025 Fall.
- *Head Teaching Assistant, Co-Instructor*. Stanford. CS 224N: Natural Language Processing. 2021, 2023.
- *Teaching Assistant*. Stanford. CS 224N: Natural Language Processing. 2020.
- *Mentor*. Stanford. General Advising, 2019-2022.
- *Mentor*. Stanford. CURIS: Summer Research for Undergraduates 2019, 2023.
- *Mentor*. Stanford. ROHU: Research Office Hours for Undergraduates.
- *Mentor*. Stanford. AI Undergraduate Mentorship Program.
- *Teaching Assistant*. Penn. CIS 530: Computational Linguistics. 2018..
- *Teaching Assistant*. Penn. CIS 121: Data Structures and Algorithms. 2015, 2016..
- *Volunteer Instructor*. Old Rochester Regional High School. The Math that Runs the World.

Students Mentored

- Tianyi Lorena Yan. *PhD*.
- Sarah Chen. *BS. Topic: Interpretability*.
- Lora Xie. *BS. Topic: Interpretability*.
- Edward Adams. *BS. Topic: Interpretability*.
- Ruth-Ann Armstrong. *MS. Topic: Jamaican Patois NLI + Multilinguality*.
- Benjamin Newman. *BS/MS. Topic: Understanding LMs*.
- Ethan Chi. *BS Topic: Multilingual probing*.

Professional Tutorials

- Generating Text from Language Models
Afra Amini, Clara Meister, John Hewitt, Luca Malagutti, Ryan Cotterell, Tiago Pimentel.
Association for Computational Linguistics (ACL) July 2023.

PATENTS

- John Hewitt, Aida Nematzadeh, and Adhiguna Kuncoro.
Determining training data sizes for training smaller neural networks using shrinking estimates.
US Patent App 18/932,554. November 2023 (Priority to Oct 2023). Assigned to DeepMind Technologies Ltd.
- John Hewitt.
Capturing Rich Response Relationships with Small-Data Neural Networks.
US Patent App 15/841,963. December 2017 (granted 2019-08-13). Assigned to Qualtrics, Inc.

PROFESSIONAL SERVICE

Organizer

- **The Learning Workshop 2026**.

Senior Area Chair

- NAACL 2025; *Interpretability Track*.

Area Chair

- COLM 2025.
- EMNLP 2023; *Interpretability Track*.
- Mechanistic Interpretability Workshop @ NeurIPS 2025.

Reviewer

- COLM 2024.
- ACL 2018 *top reviewer*, 2020 *top reviewer*, 2023.
- EMNLP 2018.
- CoNLL 2020, 2022, 2023.
- ACL Rolling Review 2021, 2022.
- Natural Language Engineering Journal 2022.
- Computational Linguistics Journal 2021.
- BlackBoxNLP 2020, 2021, 2022, 2023.
- NAACL 2021.
- EACL 2021.
- AACL 2020.
- DistShift NeurIPS Workshop on Distribution Shifts 2021, 2022.
- DeeLIO Workshop on Deep Learning Knowledge Extraction and Integration 2020.
- ACL-SRW 2019.

Departmental service

- Stanford NLP Group Social Organizer 2019–2020, 2023–2024 *ex officio*.
- Stanford CS PhD Admissions Committee 2020.

INVITED TALKS

Interplay research is alignment research with a big bet . INTERPLAY Workshop @ COLM, October 2025.

We Can't Understand AI Using our Existing Vocabulary. Seattle Minds and Machines, June 2025.

Instruction Following without Instruction Tuning. Deep Learning: Classics and Trends (ML Collective), November 2024.

Instruction Following without Instruction Tuning. Bay Area Language Processing Interest Group (Bayli), November 2024.

Instruction Following without Instruction Tuning. University of Washington, November 2024.

Instruction Following without Instruction Tuning. University of Pennsylvania, October 2024.

Understanding Language Models through Discovery and by Design. University of Michigan, February 2024.

Understanding Language Models through Discovery and by Design. Northwestern University, February 2024.

Understanding Language Models through Discovery and by Design. Harvard University, February 2024.

Understanding Language Models through Discovery and by Design. NYU, February 2024.

Understanding Language Models through Discovery and by Design. Columbia University, February 2024.

Panel on Mechanistic Interpretability. BlackBoxNLP, December 2023.

Backpack Language Models. Apple, August 2023.

Backpack Language Models. Princeton University, August 2023.

Backpack Language Models. New York University, July 2023.

Backpack Language Models. Columbia University, July 2023.

Backpack Language Models. Cornell Tech, July 2023.

Backpack Language Models. Samaya AI, June 2023.

Backpack Language Models. Anthropic, May 2023.

Backpack Language Models. Schütze Lab, LMU Munich, May 2023.

Backpack Language Models. Rycolab, ETH Zurich, April 2023.

Surviving Graduate School (panelist). ACL Mentorship Session, June 2022.

An NLP perspective on supervised analysis of neural representations. Ev Fedorenko's EvLab (MIT), December 2020.

The Unreasonable Syntactic Expressivity of RNNs. USC ISI NLP Seminar, November 2020.

Language probes as \mathcal{V} -information estimators. NLP with Friends, September 2020.

Probing Neural NLP: Ideas and Problems. Berkeley NLP Seminar, November 2019.

Emergent Linguistic Structure in Neural NLP. Amazon AI, July 2019.

A Structural Probe for Finding Syntax in Word Representations. NLP Highlights Podcast of the Allen Institute for Artificial Intelligence, May 2019.

A Structural Probe for Finding Syntax in Word Representations. Stanford Human-Centered AI Initiative Symposium, March 18, 2019.

GRANTS & AWARDS

ACL 2023 Outstanding Paper Award.

For Backpack Language Models.

R0-FoMo 2023 Runner-Up Best Paper Paper Award.

For Learning from Canonical Examples, now called Model Editing with Canonical Examples.

BlackBoxNLP 2020 Outstanding Paper Award.

For The EOS Decision and Length Extrapolation.

Two Sigma Fellowship 4th Place Prize.

2020.

NSF Graduate Research Fellowship.

(2020) In Computer Science – Natural Language Processing

EMNLP 2019 Runner-Up Best Paper Award.

For Designing and Interpreting Probes with Control Tasks