Peter Hase

peter@cs.unc.edu · peterbhase.github.io · (919) 323-0393

EDUCATION

The University of North Carolina at Chapel Hill *Fall 2019 – Present*
Fifth-year PhD student in Computer Science *Chapel Hill, NC*
Research Area: Interpretable Machine Learning | *Advisor:* Mohit Bansal

Duke University *Fall 2015 – Spring 2019*
BS in Statistical Science | Minor in Mathematics *Durham, NC*

RESEARCH INTERESTS

Interpretable machine learning, language models, AI safety, model editing, algorithmic recourse, multi-agent communication.

PUBLICATIONS

Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models
Peter Hase, Mohit Bansal, Been Kim, Asma Ghandeharioun
Preprint on arXiv. [pdf] [code]

Summarization Programs: Interpretable Abstractive Summarization with Neural Modular Trees
Swarnadeep Saha, Shiyue Zhang, Peter Hase, Mohit Bansal
ICLR 2023. [pdf] [code]

Do Language Models Have Beliefs? Methods for Detecting, Updating, and Visualizing Model Beliefs
Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, Srinivasan Iyer
EACL 2023. [pdf] [code]

GrIPS: Gradient-free, Edit-based Instruction Search for Prompting Large Language Models
Archiki Prasad, Peter Hase, Xiang Zhou, Mohit Bansal
EACL 2023. [pdf] [code]

Are Hard Examples Also Harder to Explain? A Study with Human and Model-Generated Explanations
Swarnadeep Saha, Peter Hase, Nazneen Rajani, Mohit Bansal
EMNLP 2022. [pdf] [code]

VisFIS: Visual Feature Importance Supervision with Right-for-the-Right-Reason Objectives
Zhuofan Ying,* Peter Hase,* Mohit Bansal
NeurIPS 2022. [pdf] [code]

When Can Models Learn From Explanations? A Formal Framework for Understanding the Roles of Explanation Data
Peter Hase, Mohit Bansal
ACL 2022 Workshop on Natural Language Supervision. [pdf v2] [pdf v1] [code]

Low-Cost Algorithmic Recourse for Users With Uncertain Cost Functions
Prateek Yadav, Peter Hase, Mohit Bansal
Preprint on arXiv. [pdf] [code]

Search Methods for Sufficient, Socially-Aligned Feature Importance Explanations with In-Distribution Counterfactuals
Peter Hase, Harry Xie, Mohit Bansal
NeurIPS 2021. [pdf] [code]

FastIF: Scalable Influence Functions for Efficient Model Interpretation and Debugging
Han Guo, Nazneen Fatema Rajani, Peter Hase, Mohit Bansal, Caiming Xiong
EMNLP 2021. [pdf] [code]

Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language?
Peter Hase, Shiyue Zhang, Harry Xie, Mohit Bansal
Findings of EMNLP 2020. [pdf] [code]

Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?
Peter Hase, Mohit Bansal
ACL 2020. [pdf] [code]

Interpretable Image Recognition with Hierarchical Prototypes
Peter Hase, Chaofan Chen, Oscar Li, Cynthia Rudin
AAAI-HCOMP 2019. [pdf] [code]

Shall I Compare Thee to a Machine-Written Sonnet? An Approach to Algorithmic Sonnet Generation
John Benhardt, Peter Hase, Liuyi Zhu, Cynthia Rudin
Preprint on arXiv. [pdf] [code]

AWARDS

Google PhD Fellowship (Natural Language Processing), Google *2021*
Fellowship awarded to six students globally for research in Natural Language Processing, providing up to three years of full funding

Royster PhD Fellowship, UNC Chapel Hill *2019*
University fellowship awarded to one student in the 2019 cohort of UNC-CH computer science students, providing three years of full funding

First Prize in the PoetiX Literary Turing Test, Neukom Institute, Dartmouth College *2018*
Awarded to the top submission in an open competition for algorithmic sonnet generation hosted by Dartmouth's Neukom Institute

Nomination for Undergrad TA of the Year, Dept. of Statistical Science, Duke University *2018*
One of five undergrad nominations from faculty for the department's TA of the year award

A.J. Tannenbaum Trinity Scholarship, Duke University *2015*
A full academic merit scholarship awarded to one student from Guilford County, NC

INVITED TALKS

University of Oxford *Spring 2022*
"Explainable Machine Learning in NLP: Methods and Evaluation" [slides]

NEC Laboratories Europe *Spring 2022*
"Explainable Machine Learning in NLP: Methods and Evaluation" [slides]

National Institute for Standards and Technology (NIST) *Spring 2022*
"Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?" [slides]

Allen Institute for AI *Spring 2022*
"Do Language Models Have Beliefs? Methods for Detecting, Updating, and Visualizing Model Beliefs?" [slides]

Uber AI *Spring 2022*

"The Out-of-Distribution Problem in Explainability and Search Methods for Feature Importance Explanations" [slides]

Center for Human Compatible AI (CHAI), UC Berkeley *Fall 2021*

"Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?" [slides]

Research Internships

Allen Institute for AI *Summer 2023*

Research Intern | *Supervisors:* Drs. Sarah Wiegreffe and Peter Clark *Seattle, WA*

- Studying topics at the intersection of interpretability and large language models

Google Research *Summer 2022*

Student Researcher | *Supervisors:* Drs. Asma Ghandeharioun and Been Kim *New York, NY*

- Studied methods for localizing knowledge in large language models
- Produced paper: "Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models"

Meta AI Research *Summer 2021*

Research Intern | *Supervisor:* Dr. Srinivasan Iyer *Seattle, WA*

- Studied methods for detecting and updating knowledge in language models
- Produced paper: "Do Language Models Have Beliefs? Methods for Detecting, Updating, and Visualizing Model Beliefs"

Professional Service

Program Committees *Summer 2020 – Present*

Area Chair

- ACL 2023 - Interpretability and Analysis of Models for NLP
- AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI (*Top Area Chair*)
- EMNLP 2022 - Interpretability, Interactivity and Analysis of Models for NLP

Reviewer

- NeurIPS 2023
- CVPR XAI4CV Workshop 2023
- AAAI 2023
- ACL Rolling Review, October 2022
- ACL Rolling Review, February 2022
- ACL Rolling Review, January 2022
- EMNLP 2022
- ACL Rolling Review, December 2021
- ACL Rolling Review, October 2021
- ACL Rolling Review, September 2021
- NeurIPS DistShift Workshop 2021
- EMNLP BlackboxNLP Workshop 2021
- EMNLP 2021
- ACL-IJCNLP 2021 (*Outstanding Reviewer*)
- ICLR RobustML Workshop 2021
- NAACL-HLT 2021
- EACL 2021
- EMNLP 2020 (*Outstanding Reviewer*)

TEACHING

Probabilistic Machine Learning (Graduate), Teaching Assistant *Spring 2019*
Dept. of Statistical Science, Duke University

Intro to AI, Teaching Assistant *Spring 2019*
Dept. of Computer Science, Duke University

Elements of Machine Learning, Teaching Assistant *Fall 2018*
Dept. of Computer Science, Duke University

Intro to Data Science, Teaching Assistant *Spring 2018*
Dept. of Statistical Science, Duke University

Regression Analysis, Teaching Assistant *Fall 2017*
Dept. of Statistical Science, Duke University

LEADERSHIP

Computer Science Student Association *Summer 2020 – Summer 2022*
Officer *Chapel Hill, NC*

- Organized social events for grad students including tea times, bar nights, and shared meals
- Observed faculty teaching to provide feedback in tenure review
- Recorded meeting minutes for CS faculty meetings to share with graduate students

High school and Undergraduate Research Mentoring *Spring 2020 – Present*
Research Mentor *Chapel Hill, NC*

- Meet weekly with an undergraduate research assistant in the UNC-NLP lab to support work on publication-track research
- Advised a Durham high school student on a summer project reimplementing current research in document summarization
- Presented a live research demo to Chapel Hill K-12 students for UNC CS open house

Startup Technical Advising *Fall 2019 – Fall 2021*
Technical Advisor *Chapel Hill, NC*

- curalens.ai: advised Curalens on text generation strategies for a therapeutic chat-bot (note: Curalens also advised by domain experts)
- Acta: advised Acta on approaches to automatically summarizing crowdsourced constituent feedback for efficient communication to local governments

Effective Altruism: Duke *Spring 2016 – Spring 2019*
Co-President *Durham, NC*

- Moderated weekly discussions related to Effective Altruism, the social movement centered on maximizing the good you can do for the world
- Recorded over 15 Giving What We Can pledges (10% of all future income) in pledge drives and over 30 One For the World pledges (1% of future income)
- Organized lectures and reading groups on AI safety for Duke and UNC Chapel Hill students
- Led club from 9 to 30+ active members over my tenure as Co-President