

Peter Hase

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

EDUCATION

The University of North Carolina at Chapel Hill

Fourth-year PhD student in Computer Science

Research Area: Explainable Machine Learning | Advisor: [Mohit Bansal](#)

Fall 2019 – Present

Chapel Hill, NC

Duke University

BS in Statistical Science | Minor in Mathematics

Fall 2015 – Spring 2019

Durham, NC

RESEARCH INTERESTS

Interpretable and explainable machine learning, natural language processing, multi-agent communication, AI safety.

PUBLICATIONS

VisFIS: Visual Feature Importance Supervision with Right-for-the-Right-Reason Objectives

Zhuofan Ying, Peter Hase, Mohit Bansal

Preprint on arXiv. [[pdf](#)] [[code](#)]

GrIPS: Gradient-free, Edit-based Instruction Search for Prompting Large Language Models

Archiki Prasad, Peter Hase, Xiang Zhou, Mohit Bansal

Preprint on arXiv. [[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

Preprint on arXiv. [[pdf](#)] [[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

In *NeurIPS 2021*. [[pdf](#)] [[code](#)]

When Can Models Learn From Explanations? A Formal Framework for Understanding the Roles of Explanation Data

Peter Hase, Mohit Bansal

Preprint on arXiv. [[pdf](#)] [[code](#)]

FastIF: Scalable Influence Functions for Efficient Model Interpretation and Debugging

Han Guo, Nazneen Fatema Rajani, Peter Hase, Mohit Bansal, Caiming Xiong

In *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

In *Findings of EMNLP 2020*. [[pdf](#)] [[code](#)]

Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?

Peter Hase, Mohit Bansal

In ACL 2020. [[pdf](#)] [[code](#)]

Interpretable Image Recognition with Hierarchical Prototypes

Peter Hase, Chaofan Chen, Oscar Li, Cynthia Rudin

In 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 to the Neukom Institute's open competition for algorithmic sonnet generation

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 Summer 2021

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

RESEARCH INTERNSHIPS	Google Research Student Researcher <i>Supervisors:</i> Drs. Asma Ghandeharioun and Been Kim <ul style="list-style-type: none"> Studying topics at the intersection of interpretability and large language models 	<i>Summer 2022</i> <i>New York, NY</i>
	Meta AI Research Research Intern <i>Supervisor:</i> Dr. Srinivasan Iyer <ul style="list-style-type: none"> Studied methods for detecting and updating beliefs/knowledge in language models Produced paper on the topic, “Do Language Models Have Beliefs? Methods for Detecting, Updating, and Visualizing Model Beliefs” 	<i>Summer 2021</i> <i>Seattle, WA</i>
PROFESSIONAL SERVICE	Program Committees Area Chair <ul style="list-style-type: none"> EMNLP 2022 - Interpretability, Interactivity and Analysis of Models for NLP Reviewer <ul style="list-style-type: none"> EMNLP 2022 ACL Rolling Review, February 2022 ACL Rolling Review, January 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 (<i>Outstanding Reviewer</i>) ICLR RobustML Workshop 2021 NAACL-HLT 2021 EACL 2021 EMNLP 2020 (<i>Outstanding Reviewer</i>) 	<i>Summer 2020 – Present</i>
TEACHING	Probabilistic Machine Learning (Graduate) , Teaching Assistant Dept. of Statistical Science, Duke University	<i>Spring 2019</i>
	Intro to AI , Teaching Assistant Dept. of Computer Science, Duke University	<i>Spring 2019</i>
	Elements of Machine Learning , Teaching Assistant Dept. of Computer Science, Duke University	<i>Fall 2018</i>
	Intro to Data Science , Teaching Assistant Dept. of Statistical Science, Duke University	<i>Spring 2018</i>
	Regression Analysis , Teaching Assistant Dept. of Statistical Science, Duke University	<i>Fall 2017</i>
LEADERSHIP	Computer Science Student Association Officer <ul style="list-style-type: none"> Organize 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 	<i>Summer 2020 – Present</i> <i>Chapel Hill, NC</i>

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 (met weekly)
- Presented a live research demo to Chapel Hill K-12 students for UNC CS open house

Start-up Technical Advising*Fall 2019 – Fall 2021*

Technical Advisor

Chapel Hill, NC

- curalens.ai: previously advised Curalens on text generation strategies for a therapeutic chat-bot (note: Curalens also advised by domain experts)
- [Acta](#): previously 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