

# Peter Hase

peter@cs.unc.edu • [peterbhase.github.io](https://peterbhase.github.io) • (919) 323-0393

## EDUCATION

### The University of North Carolina at Chapel Hill

PhD in Computer Science

Advisor: Dr. Mohit Bansal

Thesis: [Interpretable and Controllable Language Models](#)

Fall 2019 – May 2024

Chapel Hill, NC

### Duke University

BS in Statistical Science | Minor in Mathematics

Fall 2015 – Spring 2019

Durham, NC

## EXPERIENCE

### Schmidt Sciences

Visiting Scientist | Supervisor: Michael Belinsky

May 2025 – Present

New York, NY

### Anthropic

Resident | Supervisor: Dr. Sam Bowman

August 2024 – February 2025

San Francisco, CA

### Allen Institute for AI

Research Intern | Supervisors: Drs. Sarah Wiegrefe and Peter Clark

Summer 2023

Seattle, WA

### Google Research

Student Researcher | Supervisors: Drs. Asma Ghandeharioun and Been Kim

Summer 2022

New York, NY

### Meta FAIR

Research Intern | Supervisor: Dr. Srinivasan Iyer

Summer 2021

Seattle, WA

## RESEARCH INTERESTS

AI safety, NLP, interpretability, model editing, scalable oversight, multi-agent communication

## PUBLICATIONS

[Google Scholar](#) | citations = 2588 | h-index = 17 | i10-index = 20

### Reasoning Models Don't Always Say What They Think

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R Bowman, Jan Leike, Jared Kaplan, Ethan Perez

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

### Unlearning Sensitive Information in Multimodal LLMs: Benchmark and Attack-Defense Evaluation

Vaidehi Patil, Yi-Lin Sung, Peter Hase, Jie Peng, Tianlong Chen, Mohit Bansal

TMLR 2025. [\[pdf\]](#) [\[code\]](#)

### Teaching Models to Balance Resisting and Accepting Persuasion

Elias Stengel-Eskin, Peter Hase, Mohit Bansal

NAACL 2025. [\[pdf\]](#) [\[code\]](#)

### System-1.x: Learning to Balance Fast and Slow Planning with Language Models

Swarnadeep Saha, Archiki Prasad, Justin Chih-Yao Chen, Peter Hase, Elias Stengel-Eskin, Mohit Bansal

ICLR 2025. [\[pdf\]](#) [\[code\]](#)

**Fundamental Problems With Model Editing: How Should Rational Belief Revision Work in LLMs?**

Peter Hase, Thomas Hofweber, Xiang Zhou, Elias Stengel-Eskin, Mohit Bansal  
*TMLR 2024*. [[pdf](#)] [[code](#)]

**Unlearning Sensitive Information in Multimodal LLMs: Benchmark and Attack-Defense Evaluation**

Vaidehi Patil, Yi-Lin Sung, Peter Hase, Jie Peng, Tianlong Chen, Mohit Bansal  
*TMLR 2024*. [[pdf](#)] [[code](#)]

**LACIE: Listener-Aware Finetuning for Confidence Calibration in Large Language Models**

Elias Stengel-Eskin, Peter Hase, and Mohit Bansal  
*NeurIPS 2024*. [[pdf](#)] [[code](#)]

**Are Language Models Rational? The Case of Coherence Norms and Belief Revision**

Thomas Hofweber, Peter Hase, Elias Stengel-Eskin, and Mohit Bansal  
*Preprint on arXiv*. [[pdf](#)]

**Rethinking Machine Unlearning for Large Language Models**

Sijia Liu, Yuanshun Yao, and 11 others including Peter Hase  
*Nature Machine Intelligence*. [[pdf](#)]

**Foundational Challenges in Assuring Alignment and Safety of Large Language Models**

Usman Anwar and 37 others including Peter Hase  
*TMLR 2024*. [[pdf](#)]

**The Unreasonable Effectiveness of Easy Training Data for Hard Tasks**

Peter Hase, Mohit Bansal, Peter Clark, Sarah Wiegrefe  
*ACL 2024*. [[pdf](#)] [[code](#)]

**Can Sensitive Information Be Deleted From LLMs? Objectives for Defending Against Extraction Attacks**

Vaidehi Patel,\* Peter Hase,\* Mohit Bansal  
*ICLR 2024 (Spotlight)*. [[pdf](#)] [[code](#)]

**INSPIRE: Incorporating Diverse Feature Preferences in Recourse**

Prateek Yadav, Peter Hase, Mohit Bansal  
*TMLR 2024*. [[pdf](#)] [[code](#)]

**Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback**

Stephen Casper, Xander Davies, and 30 others including Peter Hase  
*TMLR 2023 (Outstanding Paper Finalist)*. [[pdf](#)]

**Can Language Models Teach Weaker Agents? Teacher Explanations Improve Students via Personalization**

Swarnadeep Saha, Peter Hase, Mohit Bansal  
*NeurIPS 2023*. [[pdf](#)] [[code](#)]

**Adaptive Contextual Perception: How to Generalize to New Backgrounds and Ambiguous Objects**

Zhuofan Ying, Peter Hase, Mohit Bansal  
*NeurIPS 2023*. [[pdf](#)] [[code](#)]

**Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models**

Peter Hase, Mohit Bansal, Been Kim, Asma Ghandeharioun  
*NeurIPS 2023 (Spotlight)*. [[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 (Spotlight). [\[pdf v2\]](#) [\[pdf v1\]](#) [\[code\]](#)

**The Out-of-Distribution Problem in Explainability and Search Methods for Feature Importance Explanations**

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

**Outstanding Area Chair (ACL 2023)**, Association for Computational Linguistics

2023

Recognition for metareviews for the ACL 2023 conference, “comparable in scope to the best paper awards policy (1-1.5% of the pool of reviewers and chairs)”

<b>Google PhD Fellowship (Natural Language Processing)</b> , Google	2021
Fellowship awarded to six students globally for research in Natural Language Processing, providing up to three years of full funding	
<b>Royster PhD Fellowship</b> , UNC Chapel Hill	2019
University fellowship awarded to one student in the 2019 cohort of UNC Chapel Hill computer science students, providing three years of full funding	
<b>First Prize in the PoetiX Literary Turing Test</b> , Neukom Institute, Dartmouth College	2018
Awarded to the top submission in an open competition for algorithmic sonnet generation hosted by Dartmouth's Neukom Institute	
<b>Nomination for Undergrad TA of the Year</b> , Dept. of Statistical Science, Duke University	2018
One of five undergrad nominations from faculty for the department's TA of the year award	
<b>A.J. Tannenbaum Trinity Scholarship</b> , Duke University	2015
A full academic merit scholarship awarded to one student from Guilford County, NC	

#### INVITED TALKS

<b>University of Chicago, CS and DSI Joint Colloquium</b>	Spring 2025
"AI Safety Through Interpretable and Controllable Language Models" [ <a href="#">slides</a> ]	
<b>TTIC, Young Researcher Seminar Series</b>	Fall 2024
"AI Safety Through Interpretable and Controllable Language Models" [ <a href="#">slides</a> ]	
<b>Harvard University</b>	Spring 2024
"Controlling and Editing Knowledge in Large Language Models" [ <a href="#">slides</a> ]	
<b>Pacific Northwest National Laboratories</b>	Spring 2024
"Controlling and Editing Knowledge in Large Language Models" [ <a href="#">slides</a> ]	
<b>Stanford NLP Seminar</b>	Spring 2024
"Controlling and Editing Knowledge in Large Language Models" [ <a href="#">slides</a> ]	
<b>OpenAI</b>	Spring 2024
"The Unreasonable Effectiveness of Easy Training Data for Hard Tasks" [ <a href="#">slides</a> ]	
<b>CHAI, UC Berkeley</b>	Spring 2024
"The Unreasonable Effectiveness of Easy Training Data for Hard Tasks" [ <a href="#">slides</a> ]	
<b>Brown University</b>	Spring 2023
"Interpretable and Controllable Language Models" [ <a href="#">slides</a> ]	
<b>Princeton University</b>	Spring 2023
"Interpretable and Controllable Language Models" [ <a href="#">slides</a> ]	
<b>New York University</b>	Spring 2023
"Interpretable and Controllable Language Models" [ <a href="#">slides</a> ]	
<b>University of Pennsylvania</b>	Spring 2023
"Interpretable and Controllable Language Models" [ <a href="#">slides</a> ]	
<b>University of Oxford</b>	Spring 2022
"Explainable Machine Learning in NLP: Methods and Evaluation" [ <a href="#">slides</a> ]	
<b>NEC Laboratories Europe</b>	Spring 2022
"Explainable Machine Learning in NLP: Methods and Evaluation" [ <a href="#">slides</a> ]	

**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](#)]

**CHAI, UC Berkeley***Fall 2021*

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

**ACADEMIC SERVICE****Organization**

- Workshop: Towards Knowledgeable Language Models (ACL 2024)
- Workshop: Representation Learning for NLP (ACL 2024)

**Program Committees****Senior Area Chair**

- EMNLP 2025 - Interpretability, Model Editing, Transparency, and Explainability
- ACL 2025 - Interpretability and Analysis of Models for NLP

**Area Chair**

- EMNLP 2024 - Interpretability and Analysis of Models for NLP
- EACL 2024 - Interpretability and Analysis of Models for NLP
- ACL 2023 - Interpretability and Analysis of Models for NLP  
(*Outstanding Area Chair*)
- 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**

- ICLR 2025
- NeurIPS 2024
- ACL Rolling Review, February 2024
- ICLR 2024
- AAAI 2024
- ACL Rolling Review, August 2023
- EMNLP 2023
- 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

<b>Probabilistic Machine Learning (Graduate)</b> , Teaching Assistant Dept. of Statistical Science, Duke University	<i>Spring 2019</i>
<b>Intro to AI</b> , Teaching Assistant Dept. of Computer Science, Duke University	<i>Spring 2019</i>
<b>Elements of Machine Learning</b> , Teaching Assistant Dept. of Computer Science, Duke University	<i>Fall 2018</i>
<b>Intro to Data Science</b> , Teaching Assistant Dept. of Statistical Science, Duke University	<i>Spring 2018</i>
<b>Regression Analysis</b> , Teaching Assistant Dept. of Statistical Science, Duke University	<i>Fall 2017</i>

## LEADERSHIP

<b>Research Mentoring</b>	<i>Spring 2020 – Fall 2023</i>
<ul style="list-style-type: none"> <li>• Mentored Vaidehi Patil, an early-stage PhD student, resulting in an ICLR 2024 Spotlight paper</li> <li>• Mentored Zhuofan Ying, now a PhD student at Columbia, resulting in two projects published in NeurIPS</li> <li>• Mentored Harry Xie, now a PhD student at CMU, resulting in two projects published in Findings of EMNLP and NeurIPS</li> </ul>	
<b>Wilson Center AI Policy Pipeline Program</b>	<i>Fall 2022 – Summer 2023</i>
<ul style="list-style-type: none"> <li>• Researched policy issues in explainable AI and practiced memo writing for AI policy</li> <li>• Completed policy-making training with the Wilson Center Science and Technology Innovation Program, including educational sessions with current staffers and policymakers</li> </ul>	
<b>Computer Science Student Association</b>	<i>Summer 2020 – Summer 2022</i>
Officer	<i>Chapel Hill, NC</i>
<ul style="list-style-type: none"> <li>• Organized social events for grad students including tea times, bar nights, and shared meals</li> <li>• Observed faculty teaching to provide feedback in tenure review</li> <li>• Recorded meeting minutes for CS faculty meetings to share with graduate students</li> </ul>	
<b>Startup Technical Advising</b>	<i>Fall 2019 – Fall 2021</i>
<ul style="list-style-type: none"> <li>• <a href="#">curalens.ai</a>: advised Curalens on text generation strategies for a therapeutic chat-bot (note: Curalens also advised by domain experts)</li> <li>• <a href="#">Acta</a>: advised Acta on approaches to automatically summarizing crowdsourced constituent feedback for efficient communication to local governments</li> </ul>	