Peter Hase

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

Education

The University of North Carolina at Chapel Hill — *Fall 2019 – May 2024*
PhD in Computer Science — *Chapel Hill, NC*
Advisor: Dr. Mohit Bansal
Thesis: Interpretable and Controllable Language Models

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

Experience

Stanford University — *July 2025 – Present*
Postdoctoral Researcher | *Supervisor:* Dr. Christopher Potts — *Palo Alto, CA*

Schmidt Sciences — *May 2025 – Present*
AI Institute Fellow | *Supervisor:* Michael Belinsky — *New York, NY*

Anthropic — *August 2024 – February 2025*
Resident | *Supervisor:* Dr. Sam Bowman — *San Francisco, CA*

Allen Institute for AI — *Summer 2023*
Research Intern | *Supervisors:* Drs. Sarah Wiegreffe and Peter Clark — *Seattle, WA*

Google Research — *Summer 2022*
Student Researcher | *Supervisors:* Drs. Asma Ghandeharioun and Been Kim — *New York, NY*

Meta FAIR — *Summer 2021*
Research Intern | *Supervisor:* Dr. Srinivasan Iyer — *Seattle, WA*

Research Interests

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

Fellowships

Schmidt Sciences AI Institute Fellowship — 2025
Fellowship in the amount of $300,000, covering research salary, compute, and travel

Google PhD Fellowship (Natural Language Processing) — *2021*
Research fellowship in the amount of $224,238, awarded to six students globally

Royster PhD Fellowship, UNC Chapel Hill — *2019*
Research fellowship in the amount of $250,000

Publications

Google Scholar | citations = 3396 | h-index = 20 | i10-index = 23

Unsupervised Elicitation of Language Models
Jiaxin Wen, Zachary Ankner, Arushi Somani, Peter Hase, Samuel Marks, Jacob Goldman-Wetzler, Linda Petrini, Henry Sleight, Collin Burns, He He, Shi Feng, Ethan Perez, Jan Leike
Preprint on arXiv. [pdf]

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]

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 Wiegreffe
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]

HONORS

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)"

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

ICML 2025 Workshop on Machine Unlearning for Generative AI *Summer 2025*
"Beyond Retain and Forget Sets: Unlearning as Rational Belief Revision" [slides]

University of Chicago, CS and DSI Joint Colloquium *Spring 2025*
"AI Safety Through Interpretable and Controllable Language Models" [slides]

TTIC, Young Researcher Seminar Series *Fall 2024*
"AI Safety Through Interpretable and Controllable Language Models" [slides]

Harvard University *Spring 2024*
"Controlling and Editing Knowledge in Large Language Models" [slides]

Pacific Northwest National Laboratories *Spring 2024*
"Controlling and Editing Knowledge in Large Language Models" [slides]

Stanford NLP Seminar *Spring 2024*
"Controlling and Editing Knowledge in Large Language Models" [slides]

OpenAI *Spring 2024*
"The Unreasonable Effectiveness of Easy Training Data for Hard Tasks" [slides]

CHAI, UC Berkeley *Spring 2024*
"The Unreasonable Effectiveness of Easy Training Data for Hard Tasks" [slides]

Brown University *Spring 2023*
"Interpretable and Controllable Language Models" [slides]

Princeton University *Spring 2023*
"Interpretable and Controllable Language Models" [slides]

New York University *Spring 2023*
"Interpretable and Controllable Language Models" [slides]

University of Pennsylvania *Spring 2023*
"Interpretable and Controllable Language Models" [slides]

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]

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

- NeurIPS 2025 Mechanistic Interpretability Workshop
- 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

- CHI 2025
- Widening NLP Workshop at EMNLP 2025
- 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

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

Research Mentoring *Spring 2020 – Fall 2023*

- Mentored Vaidehi Patil, an early-stage PhD student, resulting in an ICLR 2024 Spotlight paper
- Mentored Zhuofan Ying, now a PhD student at Columbia, resulting in two projects published in NeurIPS
- Mentored Harry Xie, now a PhD student at CMU, resulting in two projects published in Findings of EMNLP and NeurIPS

Wilson Center AI Policy Pipeline Program *Fall 2022 – Summer 2023*

- Researched policy issues in explainable AI and practiced memo writing for AI policy

- Completed policy-making training with the Wilson Center Science and Technology Innovation Program, including educational sessions with current staffers and policymakers

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

Startup Technical Advising *Fall 2019 – Fall 2021*

- 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