Research Statement

Peter Hase, **Anthropic**

Executive Summary

My research focuses on problems in **NLP and AI Safety**, where my goal is to develop **interpretable and controllable language models**. The areas I am most interested in include:

1. Interpretability (13 publications: 4 NeurIPS, 2 EMNLP, 1 ICLR, 1 ACL, 1 TMLR, etc.)
2. Model Editing (5 publications: 2 TMLR, 1 ICLR, 1 Nature Machine Intelligence, 1 EACL)
3. Scalable Oversight (2 publications: 1 ACL, 1 NeurIPS)

Progress in these areas will be critical for ensuring that AI systems operate in a transparent and adjustable manner, laying the foundation for safe deployments of AI in society.

Introduction and Motivation

In a 2022 survey, 36% of NLP experts agreed that "AI decisions could cause nuclear-level catastrophe" in this century [Michael et al., 2023]. This survey was conducted prior to the release of ChatGPT.

The research community's now-common concern about extreme risks from AI highlights that long-standing problems in AI safety are as important as ever. Large language models (LLMs) are increasingly being used not only for decision support but also as the backbone of autonomous AI agents. While these applications show promise in automating menial human labor and even accelerating progress in the sciences, they pose key risks of misuse and over-reliance, or in the case of agents, misalignment between their goals and human values. Work on better understanding LLM capabilities and improving model safety will be crucial for reducing these risks. To address these challenges, my research focuses on foundational problems at the intersection of NLP and AI safety:

1. **Interpretability**: It is hard to verify that ML models are *right for the right reasons.* Verifying model reasoning is crucial for ensuring proper generalization, as we cannot test models on every possible input they might face in deployment.
2. **Controllability**: We want to be able to steer individual behaviors in models on demand, since pre-trained models will need continual adjustment of specific knowledge and beliefs about the world.

Often, these topics intersect. Interpretability methods should improve our causal understanding of models and therefore our ability to control their behaviors. My research in these areas has earned spotlight awards at top machine learning conferences (NeurIPS, ICLR), been recognized through a Google PhD Fellowship, and reached wider audiences through Nature magazine [Hutson, 2024] and popular podcasts like TWIML [Charrington, 2024].

Interpretable and Controllable Language Models

Evaluating Interpretability. Historically, many interpretability methods have been introduced with convincing case studies, only to later fail basic sanity checks, like providing similar explanations for randomly initialized models as for trained models [Adebayo et al., 2018]. As a result, there has been a clear need for stronger evaluations in the area.

We developed human subject tests for rigorously assessing explanation *faithfulness*, i.e. how well explanations represent model reasoning (300+ citations: Hase and Bansal [2020]). Using *simulation* tests (Fig. 1), we showed that a number of widely popular interpretability methods failed to improve causal understanding of even small neural networks on relatively simple tasks, with the one exception of local linear explanations on tabular data. Our framework has influenced later work evaluating both mechanistic interpretability methods as well as Chain-of-Thought explanations [Anwar et al., 2024].

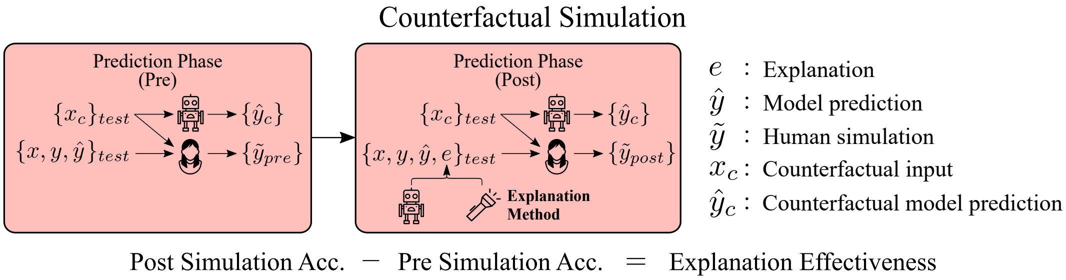
Counterfactual Simulation
Prediction Phase (Pre)
$\{x_c\}_{test}$
$\{\hat{y}_c\}$
$\{x, y, \hat{y}\}_{test}$
$\{\tilde{y}_{pre}\}$
Prediction Phase (Post)
$\{x_c\}_{test}$
$\{\hat{y}_c\}$
$\{x, y, \hat{y}, e\}_{test}$
$\{\tilde{y}_{post}\}$
Explanation Method
e : Explanation
$\hat{y}$: Model prediction
$\tilde{y}$: Human simulation
x_c : Counterfactual input
$\hat{y}_c$: Counterfactual model prediction
Post Simulation Acc. − Pre Simulation Acc. = Explanation Effectiveness

Figure 1: Simulation tests measure whether explanations help us predict model behavior on new inputs.

Natural Language Explanations. Following our previous work in interpretability, we conducted the first faithfulness evaluation of model-generated natural language explanations for models that generate explanations before their answers, like in Chain-of-Thought reasoning [Hase et al., 2020]. Experiments showed that language models with around 500M parameters could produce explanations that were genuinely faithful to their reasoning process, rather than just superficially plausible to human readers. Additionally, we were able to improve explanation faithfulness by optimizing agents for faithfulness in a multi-agent communication game. Since then, Chain-of-Thought faithfulness has proven to be a key problem in interpretability for LLMs [Turpin et al., 2023]. Later in this statement, I describe how insights from multi-agent communication could be used to improve CoT faithfulness.

Supervised Reasoning. One of the main goals of interpretability is to ensure that models are *right for the right reasons*. A key pathway to this goal is supervising model reasoning, using an interpretability method to reveal the reasoning. We introduced objectives for supervising the features that VQA models rely on in order to reduce their reliance on well-known spurious correlations common in many VQA datasets [Ying et al., 2022]. By supervising differentiable measures of feature importance in models, we were able to improve in-distribution and out-of-distribution accuracy by up to 7 points across three standard benchmarks with SOTA model architectures.

Before this, we studied how language models might learn from natural language explanations to improve their task performance [Hase and Bansal, 2021], building on predecessor methods to Chain-of-Thought prompting [Camburu et al., 2018]. We developed a synthetic task that was solvable by a model retrieving explanations seen for similar training points, but not by a model finetuned for the task, showing how explanation data could augment traditional input-output supervision. This paper received a spotlight at the 2022 ACL workshop on Learning with Natural Language Supervision.

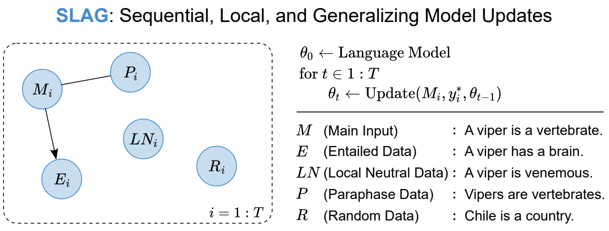
SLAG: Sequential, Local, and Generalizing Model Updates
M_i
P_i
LN_i
E_i
R_i
$i = 1 : T$
$\theta_0 \leftarrow$ Language Model
for $t \in 1 : T$
$\theta_t \leftarrow \text{Update}(M_i, y_i^*, \theta_{t-1})$
M (Main Input) : A viper is a vertebrate.
E (Entailed Data) : A viper has a brain.
LN (Local Neutral Data) : A viper is venemous.
P (Paraphase Data) : Vipers are vertebrates.
R (Random Data) : Chile is a country.

Figure 2: Framework for Model Editing.

Model Editing. In addition to better understanding models via interpretability, we want to control model behaviors in a fine-grained way. In particular, we want to edit model knowledge and beliefs whenever needed, without having to repeat any expensive pretraining or finetuning. In Hase et al. [2021], we developed a learned optimizer for editing multiple facts in an LLM sequentially (Fig. 2) while encouraging proper generalization w.r.t. (1) semantically equivalent facts (paraphrases), (2) logical consequences of updated facts (entailed facts), and (3) unrelated knowledge. Previous methods *completely failed* to edit multiple facts in a row and did not evaluate edits for their logical consequences. Since then, we have explored conceptual challenges with model editing in a collaboration with a philosopher at UNC [Hase et al., 2024b]. This is the first work to compare model editing in LLMs to rational belief revision in Bayesian agents, formalizing the model editing problem and empirically demonstrating how LLMs fall short of gold-standard Bayesian reasoning when adopting new information.

One important application of model editing is *unlearning* sensitive information from LLMs. We leveraged insights from interpretability to develop new unlearning methods (ICLR spotlight: Patil et al. [2023]). This is an important problem as pretrained LLMs are increasingly learning about dangerous areas including bioweapons and cyberattacks [Li et al., 2024], and regulation in the U.S. and Europe requires that private information can be deleted from AI systems [Henderson et al., 2023]. After we showed that we could extract information from a model that had supposedly been deleted by past SOTA methods, we proposed a new unlearning method that could lower information extraction attack success rates from 59% to 2% by scrubbing information from multiple hidden layers in the model.

Mechanistic Interpretability. The aim of mechanistic interpretability is to identify causal mechanisms underlying model behavior. In this vein, we have investigated how hidden representations influence LLM's expression of knowledge, connecting claims about interpretability to methods for controlling models (NeurIPS spotlight: Hase et al. [2023]). Many model editing papers have claimed to adjust individual model answers to questions by localizing the relevant knowledge to certain model weights. We showed that there is *no* relationship between knowledge localization results and editing method performance for several standard localization methods, editing methods, editing metrics, language models, and datasets. This finding suggests that where a fact is *currently* stored in an LLM is different from where an edited version of that fact *could be* stored. Our analysis has raised the bar for claims of causation in interpretability in subsequent work [Chang et al., 2024].

Scalable Oversight. Whereas work on model editing assumes we know what behavior we want to instill in models, *scalable oversight* concerns how to align models with our goals even when we struggle to exactly specify the desired behavior [Amodei et al., 2016]. For example, one application here is to develop AI assistants for novel scientific research even if we can only supervise them on solutions to known problems. We studied easy-to-hard generalization in LLMs in order to better understand how limited supervision can be effective in controlling model behavior on problems that we may not know the answer to [Hase et al., 2024a]. Interestingly, we found that LLMs could generalize to college-level STEM questions based only on weaker forms of supervision, like middle or high school questions, showing that it can be possible to train an LLM to solve complex tasks using only weaker forms of supervision. Concurrent work from OpenAI on weak-to-strong generalization reported similar findings [Burns et al., 2023].

Future Directions

My vision for NLP and AI Safety is one where LLMs faithfully communicate their beliefs and reasoning to us in natural language, and we use these reports to either verify that the models work as intended, or to intervene on individual failures of models in order to continually align their behavior with our goals. Achieving this vision requires advances in our ability to properly evaluate model explanation faithfulness and control individual model behaviors.

Interpretability Through Natural Language. When explaining their reasoning in words, LLMs can systematically misrepresent the reasoning underlying their behavior, even when generating explanations before their answers like in Chain-of-Thought reasoning [Turpin et al., 2023]. This is a core challenge in interpretability and controllability for LLMs, as it means that (1) verbal explanations may not reveal a model's true reasoning, and (2) when we supervise model reasoning as expressed by verbal explanations, this may not actually control the model's true reasoning. In this area, I am interested in research on:

1. *Improving Explanation Faithfulness*: Methods like CoT and RLHF have enabled models to produce natural language explanations alongside their solutions to tasks. In order to improve the faithfulness of such explanations, we must improve the explanations' precision and recall over causal factors that influence model decisions. If the model explanation mentions a factor, it should actually be causal (precision). If a factor is important, the model explanation should mention it (recall). New training objectives will be critical for improving LLMs' ability to verbalize these

factors in their stated reasoning. We should also leverage insights from pragmatics to improve explanations: models should highlight factors that humans expect to see or that could be surprising, and this means using theory-of-mind to tailor explanations to the recipient.

2. *Adversarial Evaluations*: The field needs challenging benchmarks for explanation faithfulness in order to hold interpretability methods to a high bar. I believe there is room for improvement in adversarial evaluation of model explanations, i.e. evaluations specifically aimed to reveal interpretability failures of models. When do models generate explanations that are inconsistent with their behavior or their explanations for other data points? Over normal input distributions, it may be hard to demonstrate such unfaithfulness. Adversarially searching through the input space should help detect failure cases. Such an approach will be valuable for preventing faithfulness evaluations from quickly saturating and reflecting over-optimism in models ability to produce consistent explanations of their behavior.

Updating Model Beliefs and Behaviors. How can we adjust individual beliefs and behaviors in LLMs when they act in undesired ways? This problem has become the focus of model editing research. Though it is a promising and growing direction, this area has run directly into old problems in continual learning, as well as deeper problems in philosophy concerning belief revision in rational agents [Hase et al., 2024b]. My goal is to make concrete progress on important applications of model editing while simultaneously building out the science of belief revision for LLM-based AI systems. I aim to explore:

1. *Unlearning Dangerous Capabilities*: As LLM capabilities continue to improve, they may soon be leveraged for malicious purposes, including bioweapon or cyberattack development [Li et al., 2024]. To prevent such risks, we need new methods for unlearning dangerous capabilities that are robust against adversarial attacks, as we can expect that deployed LLMs will face adversarial attempts to elicit such capabilities. Our previous work suggests we can leverage interpretability techniques to safeguard against whitebox attacks [Patil et al., 2023]. I believe we can further improve robustness against finetuning attacks and prompt attacks with metalearning methods that train a model to be robust to perturbations to its weight or input space. The unlearning problem also invokes important questions about what it means to fully unlearn something, particularly when models retain related knowledge that could be useful for recovering the unlearned information.
2. *Science of Beliefs in AI*: When can we explain behavior of LLMs in terms of “beliefs” and “desires”? While such a framework could help us explain LLM behavior, determining if it applies in the first place requires a better understanding of rational behavior in these systems. Key questions in this area include: How do models represent truthfulness, and do they aim to produce logically consistent statements? Do they have internal world models that enable the logical propagation of new beliefs? And do they rationally pursue goals on the basis of these beliefs? Precisely answering these questions will be crucial for making rigorous claims about LLMs and their likeness to rational agents that pursue goals in the world.

The success of this research will produce LLMs that faithfully communicate their reasoning to people, allowing us to verify whether their reasoning is trustworthy and aligned with human values. With a deeper understanding of LLMs, we will assess if they can represent and interact with the world like rational agents. Progress in these directions will address key challenges in AI safety, clarifying how we can deploy AI systems responsibly and to the benefit of humanity.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *NeurIPS*, 2018. URL https://arxiv.org/abs/1810.03292.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016. URL https://arxiv.org/pdf/1606.06565.pdf.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin Edelman, Zhaowei Zhang, Mario Gunther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong, Seán Ó hÉigeartaigh, Gabriel Rachet, Giulio Corsi, Alan Chan, Markus Anderljung, Lillian Edwards, Yoshua Bengio, Danqi Chen, Samuel Albanie, Tegan Maharaj, Jakob Foerster, Florian Tramer, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. Foundational challenges in assuring alignment and safety of large language models, 2024. URL https://arxiv.org/pdf/2404.09932.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023. URL https://cdn.openai.com/papers/weak-to-strong-generalization.pdf.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. In *NeurIPS 2018*, 2018. URL https://arxiv.org/pdf/1812.01193.pdf.

Ting-Yun Chang, Jesse Thomason, and Robin Jia. Do localization methods actually localize memorized data in llms? a tale of two benchmarks. In *NAACL-HLT*, pages 3190–3211, 2024. URL https://arxiv.org/pdf/2311.09060.

Sam Charrington. Localizing and editing knowledge in LLMs with peter hase. TWIML AI Podcast, 2024. URL https://twimlai.com/podcast/twimlai/guest/peter-hase/.

Peter Hase and Mohit Bansal. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior? In *ACL*, 2020. URL https://arxiv.org/pdf/2005.01831.pdf.

Peter Hase and Mohit Bansal. When can models learn from explanations? a formal framework for understanding the roles of explanation data. In *ACL Workshop on Learning From Natural Language Supervision*, 2021. URL https://arxiv.org/pdf/2102.02201.pdf.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In *Findings of EMNLP*, 2020. URL https://arxiv.org/abs/2010.04119.

Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. Do language models have beliefs? methods for detecting, updating, and visualizing model beliefs. In *EACL*, 2021. URL https://arxiv.org/pdf/2111.13654.pdf.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *NeurIPS*, 2023. URL https://arxiv.org/pdf/2301.04213.pdf.

Peter Hase, Mohit Bansal, Peter Clark, and Sarah Wiegreffe. The unreasonable effectiveness of easy training data for hard tasks. In *ACL*, 2024a. URL https://arxiv.org/pdf/2401.06751.

Peter Hase, Thomas Hofweber, Xiang Zhou, Elias Stengel-Eskin, and Mohit Bansal. Fundamental problems with model editing: How should rational belief revision work in llms? *TMLR*, 2024b. URL https://arxiv.org/pdf/2406.19354.

Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang. Foundation models and fair use, 2023. URL https://arxiv.org/pdf/2303.15715.pdf.

Matthew Hutson. How does chatgpt'think'? psychology and neuroscience crack open ai large language models. *Nature*, 629(8014):986–988, 2024. URL https://www.nature.com/articles/d41586-024-01314-y.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Zhenqi Zhao, Ariel Herbert-Voss, Cort B. Breuer, Samuel Marks, Oam Patel, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Liu, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The wmdp benchmark: Measuring and reducing malicious use with unlearning, 2024. URL https://arxiv.org/abs/2403.03218.

Julian Michael, Ari Holtzman, Alicia Parrish, Aaron Mueller, Alex Wang, Angelica Chen, Divyam Madaan, Nikita Nangia, Richard Yuanzhe Pang, Jason Phang, and Samuel R. Bowman. What do NLP researchers believe? results of the NLP community metasurvey. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16334–16368, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.903. URL https://aclanthology.org/2023.acl-long.903.

Vaidehi Patil, Peter Hase, and Mohit Bansal. Can sensitive information be deleted from llms? objectives for defending against extraction attacks. In *ICLR*, 2023. URL https://arxiv.org/pdf/2309.17410.pdf.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ed3fea9033a80fea1376299fa7863f4a-Paper-Conference.pdf.

Zhuofan Ying, Peter Hase, and Mohit Bansal. Visfis: Visual feature importance supervision with right-for-the-right-reason objectives. *Advances in Neural Information Processing Systems*, 35:17057–17072, 2022. URL https://arxiv.org/pdf/2206.11212.pdf.

Executive Summary

I have always deeply enjoyed teaching and mentoring others, and I feel that teaching is especially important now as so many students flock to computer science and AI. I have several years TA experience, including up to two courses per semester, and past mentees of mine are now in PhD programs after finishing their undergraduate degrees. I'm honored that two of them wrote letters of support for my faculty applications, attesting to our time working together. My goal in teaching is to equip students with deep conceptual understanding of the topics at hand and to improve their overall technical literacy.

I am particularly excited to teach introductory courses that draw students into the discipline, as well as advanced seminars closer to my areas of research expertise:

- Intro to NLP
- Intro to Data Science
- Intro to Machine Learning
- Probability
- AI Safety / AI Ethics / Responsible AI
- Interpretable Machine Learning

Teaching and Mentoring

Teaching. At UNC, all PhD students take a course on pedagogy where we practice technical lecturing skills and develop active learning exercises. The department chair commended my practice lecture on Bayes Rule for its clarity and engaging exercises. Before UNC, I gained two years experience as a teaching assistant during my undergrad at Duke. During this time, I was nominated for TA of the Year, one of five nominations in the department. My classes included Regression Analysis, Intro to Data Science, Intro to Machine Learning, Intro to AI, and Probabilistic Machine Learning (a graduate course). All of these classes involved regular office hours as well as weekly labs where I covered relevant material and exercises, often including new content, with a group of up to 25 students. I always felt that I did not truly understand a topic until I could teach it to someone else, so I believe I learned as much TAing as I did in many of my own classes.

In a faculty role, I would be especially interested in teaching introductory classes on NLP and related topics, as well as graduate level seminars on research areas including interpretable machine learning, multi-agent communication, AI ethics, and AI safety.

Mentoring. I find the idea of mentoring PhD students to be one of the most compelling features of academic work. I have previously mentored high school students (David Liu), domestic and international undergraduate students (Zhuofan Ying, Harry Xie), and early PhD students (Vaidehi Patil). Harry Xie and I worked together on projects in natural language explanation methods and feature attribution methods, which were published in Findings of EMNLP and NeurIPS, respectively. Harry graduated from Duke University and worked at Google before beginning his PhD in Statistics at Carnegie Mellon University. Zhuofan and I worked together on two projects, on supervising feature importance in VQA models and explaining foreground vs. background reliance in vision models, both of which were published at NeurIPS. Zhuofan graduated from UNC Chapel Hill and has started his PhD in Computational Neuroscience at Columbia University. Most recently, I worked with Vaidehi on a project on unlearning sensitive information from language models, which was published as a spotlight at ICLR 2024. Vaidehi continues her research in Mohit Bansal's lab at UNC. All of these mentoring relationships have been immensely fulfilling to me, and I truly enjoyed helping these students grow into independent and capable researchers in their own right. See my application materials for two letters of support that Zhuofan and Harry wrote for me about our time working together.

Science Communication

Invited Talks. During my PhD, I had the honor of giving 15 invited talks at industry, university, non-profit, and government groups, including venues such as Uber and OpenAI, Berkeley and Stanford, the Allen Institute for AI, and the National Institute for Standards and Technology (NIST). I believe it is imperative to share the results of our work widely among industry practitioners and government agencies, in order to help translate basic science into reliable AI technologies and effective regulatory policies in society. For example, NIST was interested in work we had done on evaluating explainability methods in AI, while at OpenAI I covered work on generalizing from easier data to harder data that is more costly to label.

Public Communication. AI systems have rapidly become part of everyday life, with people regularly relying on them for work and personal decision-making: ChatGPT has over 10 million weekly users. It is crucial, therefore, that the public hear from experts about how to understand these systems and what to expect from them. Previously, I have engaged with journalists, news outlets, and podcasters to reach a broader audience on AI issues. For example, our work has appeared in Nature magazine, I've been on local news in North Carolina, and I've gone on podcasts like TWIML to talk about our research. Recently, I consulted for a United Nations report on LLMs and international security.

Diversity, Equity, and Inclusion

DEI in Research. Some draw a distinction between "core" machine learning and work on fairness, value-alignment, and socio-technical work in machine learning. I reject such distinctions, since it is a critical part of academia's mission to promote the production of knowledge in a way that advances human values, and not just to cater to a field's latest intellectual interests. I aim to uphold values of diversity and inclusion as I work with others on technical AI research, mentor students, and serve my academic community. I believe that NLP and AI Safety are key areas for connecting DEI to research, as we aim to make models more interpretable to humans and align them to human values.

Community Building. I have worked to create inclusive academic spaces at both local and field-wide levels. For two years I was an officer on the Computer Science Student Association at UNC. As a part of this role, I helped organize social events for grad students including tea times, bar nights, and shared meals. We also summarized faculty meetings to keep graduate students informed about the department. This past year, I helped organize RepL4NLP, a long-running workshop to be hosted at ACL 2024, as well as the Towards Knowledgeable Language Models workshop, also at ACL 2024. My role in organizing involved managing review assignment and collection, discussing program content with the organizing committee, and recommending speakers to invite.

Paper #1

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

Authors: Peter Hase and Mohit Bansal, UNC Chapel Hill

Link: link to paper

Venue: ACL 2020 (300+ citations)

Significance: Historically, many interpretability methods have been introduced with convincing case studies, only to later fail basic sanity checks, like providing similar explanations for randomly initialized models as for trained models. As a result, there has been a clear need for stronger evaluations in the area.

We developed human subject tests for rigorously assessing explanation *faithfulness*, i.e. how well explanations represent model reasoning. Using *simulation* tests, we showed that a number of widely popular interpretability methods failed to improve causal understanding of even small neural networks on relatively simple tasks. A simulation test works by showing users model explanations for some data points, and then testing if they can predict model outputs for similar data points. This test measures the accuracy of a user's mental model of an AI system. If the explanation tells the user, in a generalizable way, how the model relies on input features in order to arrive at its predictions, they should be able to predict its outputs on other data. While our study provided strong evidence that many methods were not living up to their claims of improving model understanding, the one exception was local linear explanations on tabular data, e.g. LIME and SHAP. This one positive result helped confirm to us that interpretability for deep learning models was possible, but it was going to be a challenge for the field to solve over the coming years. Our framework has influenced later work evaluating both mechanistic interpretability methods as well as natural language explanations.

Paper #2

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

Authors: Peter Hase[1], Mohit Bansal[1], Been Kim[2], and Asma Ghadenharioun[2]

[1]UNC Chapel Hill
[2]Google DeepMind

Link: link to paper

Venue: NeurIPS 2023 (*Spotlight*; 100+ citations)

Significance: The aim of mechanistic interpretability is to identify causal mechanisms underlying model behavior. In this vein, we have investigated how hidden representations influence LLM's expression of knowledge, connecting claims about interpretability to methods for controlling models. Many model editing papers have claimed to adjust individual model answers to questions by localizing the relevant knowledge to certain model weights. We showed that there is *no* relationship between knowledge localization results and editing method performance for several standard localization methods, editing methods, editing metrics, language models, and datasets. This finding suggests that where a fact is *currently* stored in an LLM is different from where an edited version of that fact *could be* stored. Our analysis has raised the bar for claims of causation in interpretability in subsequent work and spurred a variety of follow-up investigations into what kinds of information and capabilities can be uniquely attributed to specific model components.

Paper #3

Title: Foundational Challenges in Assuring Alignment and Safety of Large Language Models

Authors: Usman Anwar, Abulhair Saparov*, Javier Rando*, Daniel Paleka*, Miles Turpin*, **Peter Hase***, Ekdeep Singh Lubana*, Erik Jenner*, Stephen Casper*, Oliver Sourbut*, Benjamin L. Edelman*, Zhaowei Zhang*, Mario Günther*, Anton Korinek*, Jose Hernandez-Orallo*, Lewis Hammond, Eric Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong, Seán Ó hÉigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Aleksandar Petrov, Christian Schroeder de Witt, Sumeet Ramesh Motwan, Yoshua Bengio, Danqi Chen, Philip H.S. Torr, Samuel Albanie, Tegan Maharaj, Jakob Foerster, Florian Tramer, He He, Atoosa Kasirzadeh, Yejin Choi, David Krueger

*Major Contribution

See paper link for author affiliations.

Link: link to paper

Venue: TMLR 2024 (95+ citations)

Significance: This paper is a 175 page agenda paper on open problems in scientific understanding of LLMs, safe development and deployment of models, and associated sociotechnical challenges. Along with Miles Turpin, I developed the section on Tools for Interpreting or Explaining Model Behavior (8 pages of the agenda). Our section covers key problems with mechanistic interpretability, natural language explanations, and formal program synthesis methods. We ultimately recommend 17 specific research questions for the community to work on. I believe that agenda papers like this are important for guiding the broader field of AI Safety. For example, as sparse-autoencoders have captured the attention of the interpretability community, many new papers have focused on small changes to objectives for training the autoencoders. We want to refocus future work on important challenges that *prevent SAEs from being useful in practice*, including that SAEs are not yet useful for controlling model behavior better than simple instruction-following, and most evaluations can be described as "searching under the streetlight" (i.e., not focused on uncovering unknown problematic model behaviors). We also make direct recommendations for work on natural language explanations – the single most common medium researchers use to understanding and diagnose problems with model reasoning – and program synthesis, an approach which could potentially offer formal guarantees for interpretability. Besides steering the field, this work makes concrete a set of research questions I would personally be interested in investigating in my own future work.