Interpretable and Controllable Language Models

UNC NLP

Peter Hase
peter@cs.unc.edu

Collaborators

First Authors: Swarnadeep Saha, Zhuofan Ying
Middle Authors: Shiyue Zhang, Harry Xie, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov
Last Authors: Mohit Bansal, Srinivasan Iyer

Quick Summary

- Interpretability
  - Motivation: **Interpretability is useful because of limitations with *testing* models**
  - Result: Natural language **explanations near human level but decline in quality for hard data**
  - Paper: "Are Hard Examples also Harder to Explain? A Study with Human and Model-Generated Explanations" (Saha et al., 2022)
- Fine-grained Control
  - Motivation: **We want to fix individual model errors** (both factual and moral) over time
  - Result: **Model editing is increasingly useful for fine-grained control but has a long way to go**
  - Paper: "Do Language Models Have Beliefs? Methods for Detecting, Updating, and Visualizing Model Beliefs" (Hase et al., 2021)
- Controllable Reasoning
  - Motivation: **If we verify the reasoning process, we don't have to exhaustively test all inputs**
  - Result: **Supervising model explanations helps with OOD generalization**
  - Paper: "VisFIS: Visual Feature Importance Supervision with Right-for-the-Right-Reason Objectives" (Ying et al., 2022)

Roadmap

Goal	Approach
Interpretability	Natural Language Explanations
Fine-grained Control	Model Editing
Controllable Reasoning	Supervising Model Explanations

Roadmap

Goal	Approach
Interpretability	Natural Language Explanations
Fine-grained Control	Model Editing
Controllable Reasoning	Supervising Model Explanations

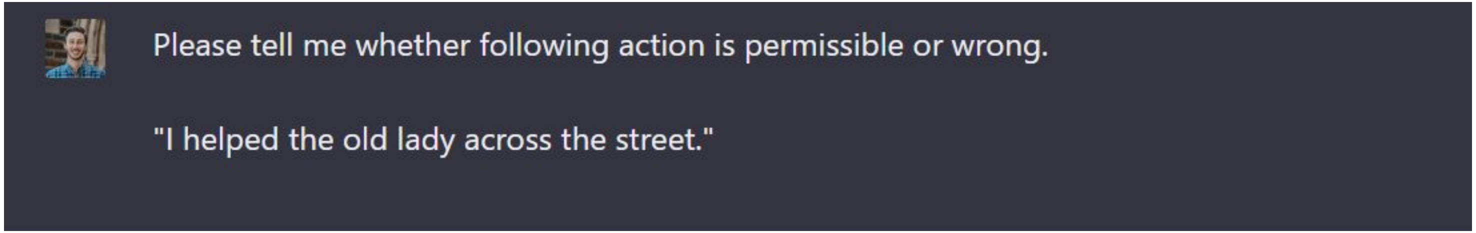
UNC NLP

Definitions

- A model is *interpretable* if we can form accurate beliefs about how it works
- “How it works” = causal chains of events that lead to model outputs

Hase et al.

What Is *Uniquely* Useful About Interpretability?

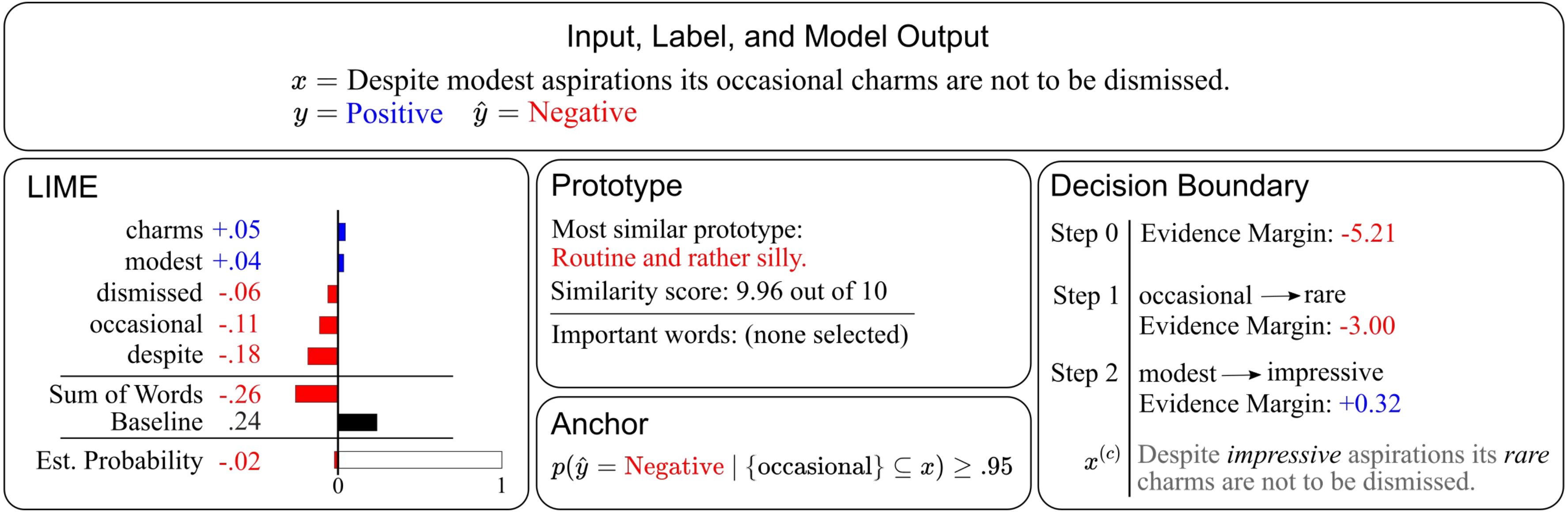


Hypothesis: “Model says action is wrong if there is risk of personal harm”

Three ways to verify this:

1. Prove it formally
2. Test it empirically (get contrastive examples to test model on)
3. Use explanation method to verify the hypothesis

What Is *Uniquely* Useful About Interpretability?



(Hase and Bansal, 2020)

UNC NLP

What Is *Uniquely* Useful About Interpretability?

Please tell me whether following action is permissible or wrong.

"I helped the old lady across the street."

It is generally considered to be a kind and helpful action to assist an elderly person or someone who may have difficulty crossing the street. It is a thoughtful gesture that can help to ensure the safety and well-being of others, and it is generally considered to be a good thing to do. There may be certain situations in which it is not advisable or appropriate to assist someone in crossing the street, such as if there is a risk of danger to oneself or the other person, but in most cases, helping an elderly person or someone in need to cross the street is a kind and helpful action.

12/28/22

What Is *Uniquely* Useful About Interpretability?

Hypothesis: "Model says action is wrong if there is risk of personal harm"

Explanation: "not advisable…if there is risk of danger to oneself or the other person"

If the explanation accurately reveals causal chain behind model behavior

Then, we're done!

Could be hard to show

Three ways to verify our hypothesis:

1. Prove it formally
2. Test it empirically (get contrastive examples to test model on)
3. *Use explanation method to verify the hypothesis* (**could be easiest method**)

UNC NLP

What Is *Uniquely* Useful About Interpretability?

Interpretability could also improve empirical testing

Hypothesis: “Model says action is wrong if there is risk of personal harm”

…if the actor has bad intentions

…???

Explanation methods could help us discover things we should be testing

Generate and Verify Hypotheses

Roadmap

Goal	Approach
Interpretability	**Natural Language Explanations**
Fine-grained Control	Model Editing
Controllable Reasoning	Supervising Model Explanations

Natural Language Explanations - Background

- Do models produce human-level free-text explanations for textual reasoning tasks?

SNLI (Bowman et al., 2015)

Premise: Dark-haired man wearing a watch and oven mitt about to cook some meat in the kitchen.
Hypothesis: A man is cooking something to eat.
Label: entailment

e-SNLI (Camburu et al., 2018): Meat is cooked in a kitchen, and is a food that you eat. Using an oven mitt implies you're about to cook with hot utensils.
GPT-3: Cooking is usually done to prepare food to eat.

	Preferred Explanation (%)		
Dataset	**Crowd**	**Tie**	**GPT-3**
CoS-E	7.2	13.9	**78.9**
ECQA	44.5	9.7	**45.7**
e-SNLI	**49.6**	9.7	40.7

"Which of two explanations best explains the answer?"

(Wiegreffe et al., 2022)

Natural Language Explanations - Background

Plausibility: Does the explanation sound like it could be **valid reasoning**?
(Jacovi and Goldberg, 2020)

- Supports label
- Generalizable reasoning pattern (not ad hoc)

Faithfulness: Does the explanation accurately represent **how the model works**?
"how model works" = causal chains of events that lead to model outputs

Plausibility on its own is dangerous

- Model capability to keep track of
- Precondition for faithfulness

Natural Language Explanations - Saha et al., 2022

"Are Hard Examples also Harder to Explain? A Study with Human and Model-Generated Explanations"
Swarnadeep Saha, Peter Hase, and Mohit Bansal. 2022. EMNLP

- Do models produce human-level free-text explanations for textual reasoning tasks? (in terms of plausibility)

Do models explain ***hard*** data as well as ***easy*** data?

- Hardness measured with minimum-description length metric (Swayamdipta et al., 2020)
- "How long does it take to learn the datapoint?" (for a finetuned model)

Natural Language Explanations - Saha et al., 2022

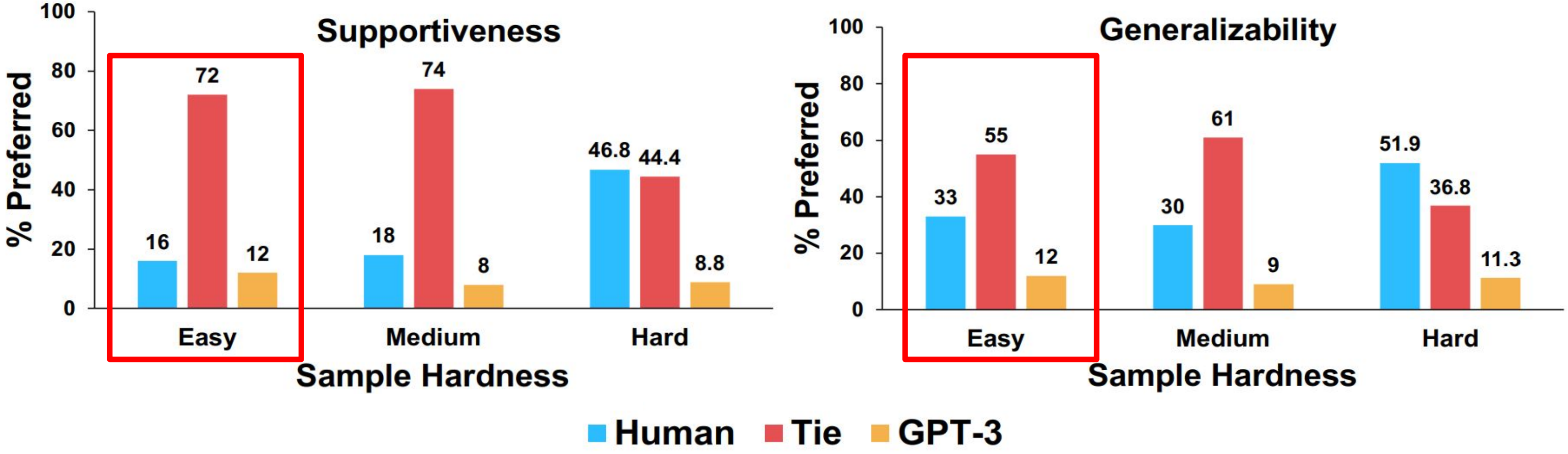
- Will measure…
  - Grammaticality
  - Label-supportiveness
  - Generalizability
- Using…
  - text-davinci-002
  - k-shot prompting with retrieval of similar data
  - MTurk for human eval
- On…
  - WinoGrande data
  - 100 points for each of three hardness groups

Sentence	Options
I wanted to buy small tweezer to fit in my wristlet, but they still didn't fit. The _ were too small.	tweezer / wristlet
The documents contained in the files could not fit properly. The _ were too large.	documents / files
I measured the area in my kitchen, but the stove didn't fit because the _ was too small.	kitchen / stove

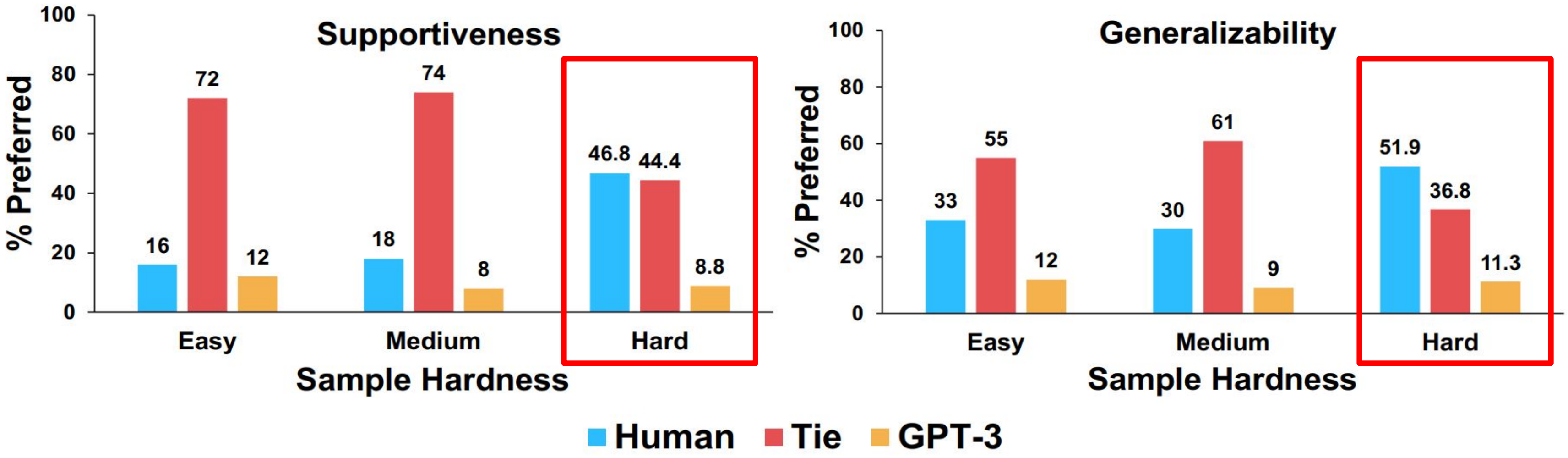
Human reasoning pattern: "**If X is larger than Y, then X does not fit in Y.**"

UNC NLP

Natural Language Explanations - Saha et al., 2022



Natural Language Explanations - Saha et al., 2022



Natural Language Explanations - Conclusion

Natural language explanations near human level but **decline in quality for hard data**
…with text-davinci-002 on WinoGrande, according to MTurkers, etc.

Roadmap

Goal	Approach
Interpretability	Natural Language Explanations
Fine-grained Control	Model Editing
Controllable Reasoning	Supervising Model Explanations

UNC NLP

Definitions

- A model is *controllable* if we can specify certain outputs for certain inputs
  - Specify *formally*: want probability of Y to be P, subject to some constraints…
  - Specify *informally*: want model to never output content that harms its readers
- *Fine-grained* control: we want to fix individual errors as we find them

UNC NLP

What Is *Uniquely* Useful About Controllability?

What awards did Mary Lowe Good receive?

Mary Lowe Good was a renowned chemist who made significant contributions to the field of organic chemistry. She received several awards throughout her career, including:

Wrong 1. Garvan Medal (1954) - This is awarded by the American Chemical Society (ACS) to recognize outstanding contributions to the field of chemistry by women.

Wrong 2. Francis P. Garvan-John M. Olin Medal (1963) - This is also awarded by the ACS, to recognize distinguished service to chemistry by women.

Wrong 3. Priestley Medal (1983) - This is the highest honor awarded by the ACS, and it recognizes distinguished services to chemistry.

Wrong 4. National Medal of Science (1991) - This is the highest scientific honor awarded by the United States government.

https://en.wikipedia.org/wiki/Mary_L._Good

What Is *Uniquely* Useful About Controllability?

- Usual argument goes like…
- Pretraining and finetuning large models over lots of data is expensive
- We can identify errors, but want to avoid re-training
- Want to fix errors one at a time

What Is *Uniquely* Useful About Controllability?

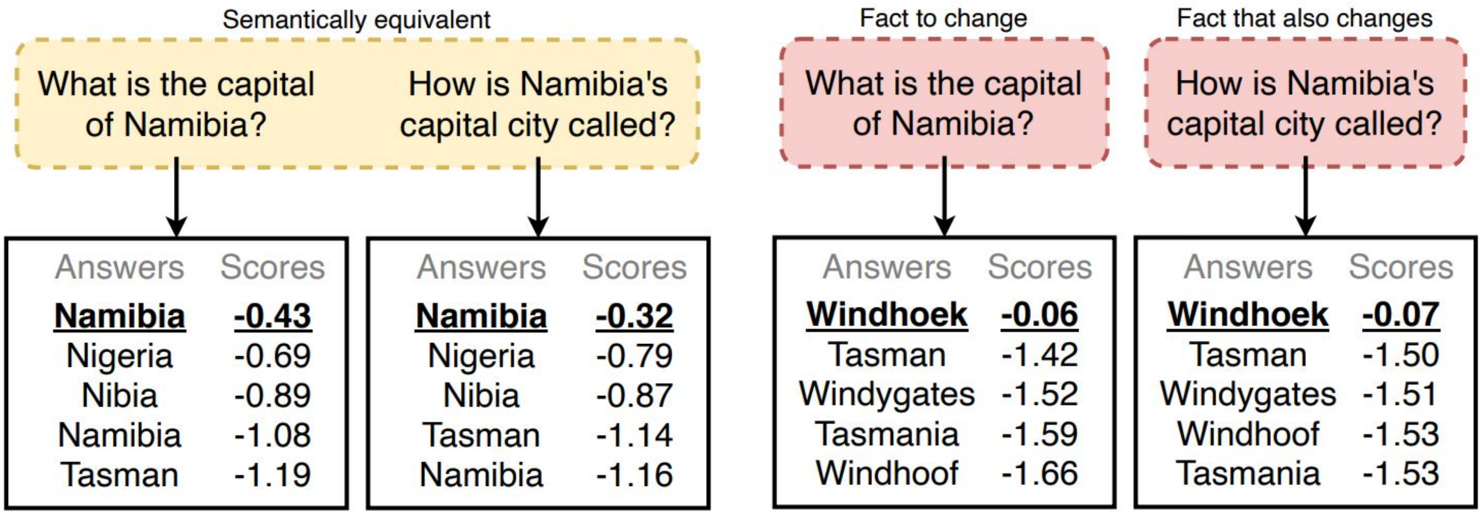
- Usual argument goes like…
- Pretraining and finetuning large models over lots of data is expensive
- We can identify errors, but want to avoid re-training - **would this even work?**
- Want to fix errors one at a time

Roadmap

Goal	Approach
Interpretability	Natural Language Explanations
Fine-grained Control	**Model Editing**
Controllable Reasoning	Supervising Model Explanations

UNC NLP

Model Editing - Background



Before Edit **After Edit**

(De Cao et al., 2020)

Model Editing - Background

- A word on terminology
- Editing = updating = revising
- What are we editing?
- “Fact” and “knowledge” seem awkward if information isn’t true
- “Belief” feels appropriately *weaker*
- Dennett (1995) characterizes a *belief* as:

An informational state decoupled from any motivational state

- This problem has been called *belief revision* in CS+philosophy since 1979 (Doyle)

UNC NLP

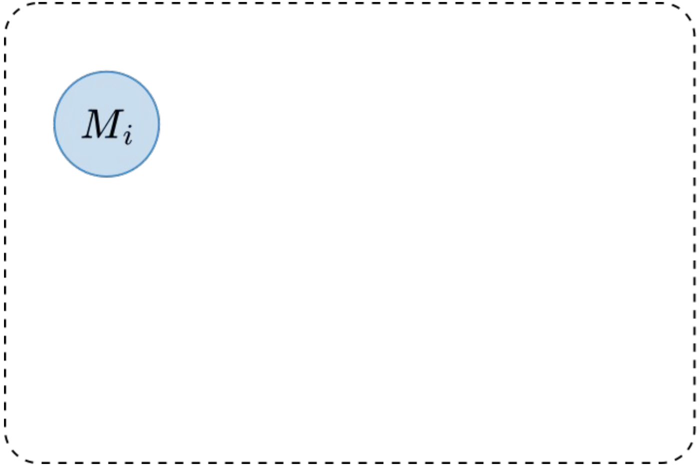
Model Editing - Hase et al., 2021

"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, and Srinivasan Iyer. 2021. EACL

- A few main research questions:
  1. How should we evaluate model edits?
  2. Can we continually update a model with new beliefs?

Model Editing - Hase et al., 2021

- How should we evaluate model edits?

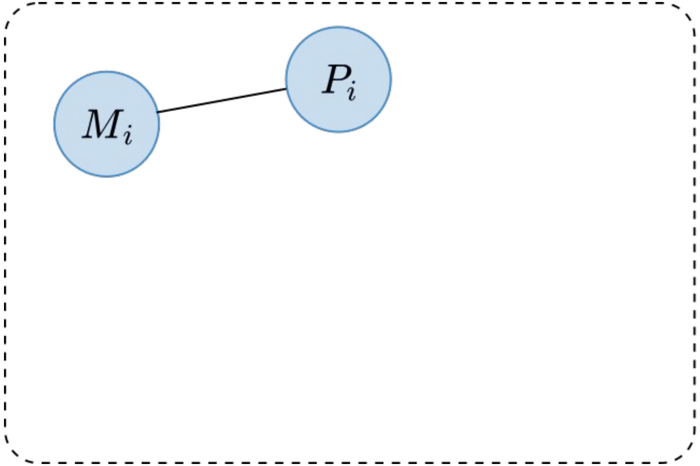


M (Main Input) : A viper is a vertebrate.

Vipers are vertebrates.

Model Editing - Hase et al., 2021

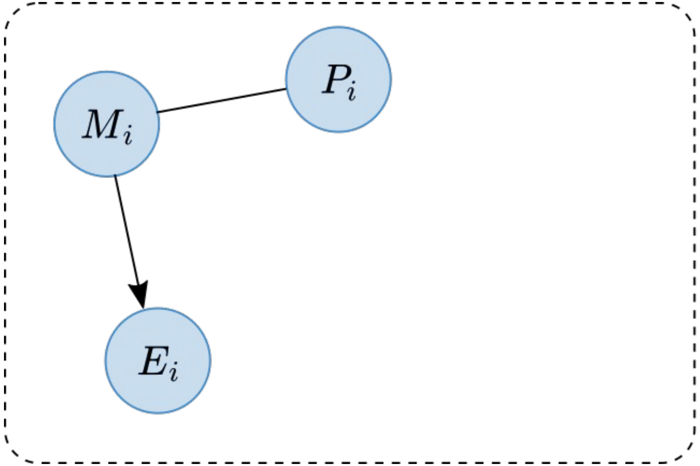
- How should we evaluate model edits?



M	(Main Input)	:	A viper is a vertebrate.
P	(Paraphase Data)	:	Vipers are vertebrates.

Model Editing - Hase et al., 2021

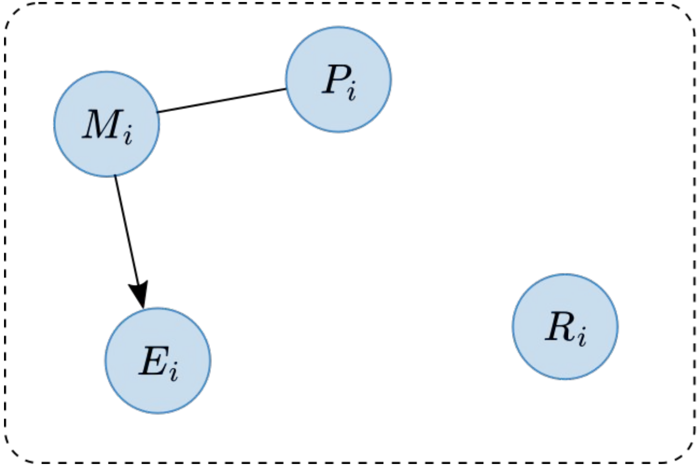
- How should we evaluate model edits?



M	(Main Input)	:	A viper is a vertebrate.
P	(Paraphase Data)	:	Vipers are vertebrates.
E	(Entailed Data)	:	A viper has a brain.

Model Editing - Hase et al., 2021

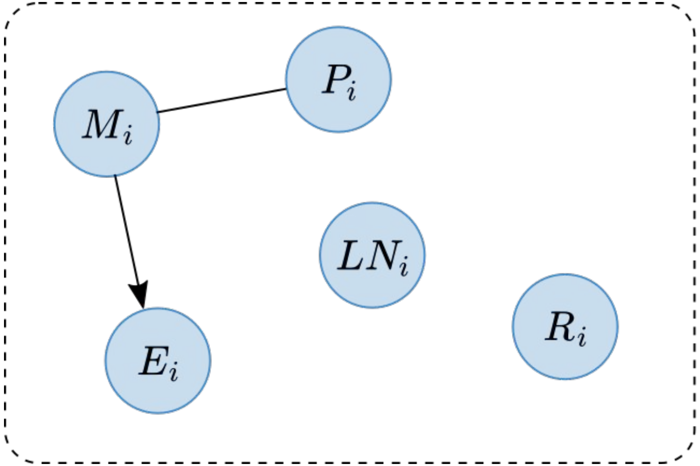
- How should we evaluate model edits?



M	(Main Input)	:	A viper is a vertebrate.
P	(Paraphase Data)	:	Vipers are vertebrates.
E	(Entailed Data)	:	A viper has a brain.
R	(Random Data)	:	Chile is a country.

Model Editing - Hase et al., 2021

- How should we evaluate model edits?



M	(Main Input)	:	A viper is a vertebrate.
P	(Paraphase Data)	:	Vipers are vertebrates.
E	(Entailed Data)	:	A viper has a brain.
R	(Random Data)	:	Chile is a country.
LN	(Local Neutral Data)	:	A viper is venemous.

Model Editing - Hase et al., 2021

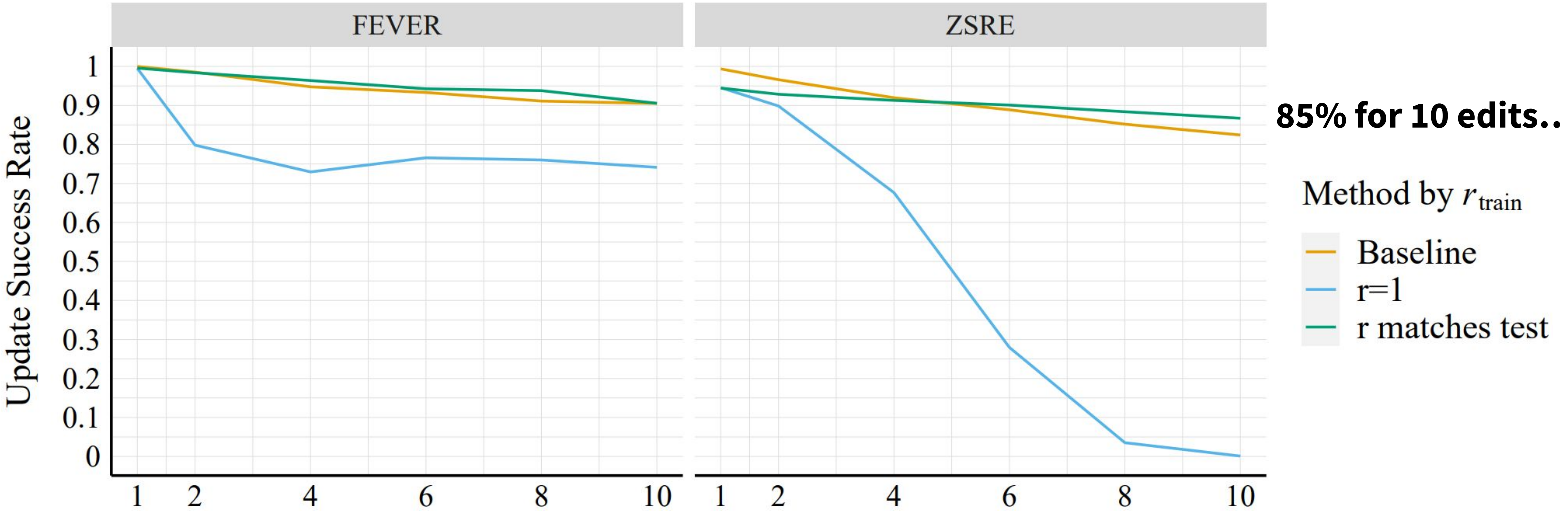
- So what are the methods and how well do they work?
- Can we continually update a model with new beliefs?
- Methods:
    1. Edit model weights
    2. Persistent memory + retrieval

UNC NLP

Model Editing - Hase et al., 2021

- Continual belief updating - **hypernetwork weight editing** on **t5-base**

Update Success (Main Input)



 # Model Edits

85% for 10 edits...

Model Editing - Hase et al., 2021

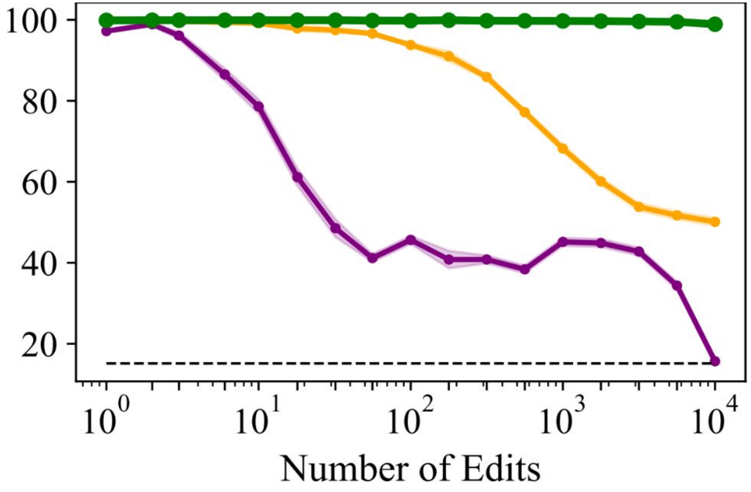
- Continual belief updating - **hypernetwork weight editing** on **t5-base**
- Want to discuss what’s happened since 2021
- But first:

1. Harder to ***fix errors*** than to ***create them***
2. Harder to retain performance on ***local data*** than ***random data***
3. Harder to generalize to ***entailed data*** than ***paraphrases***
4. Updates greatly ***improve consistency*** (model was wrong in inconsistent ways)

Model Editing - Recent Work

- Continual belief updating - **MEMIT weight editing** on **GPT-J** (Meng et al., 2022)

~= Update Success (Main Input)

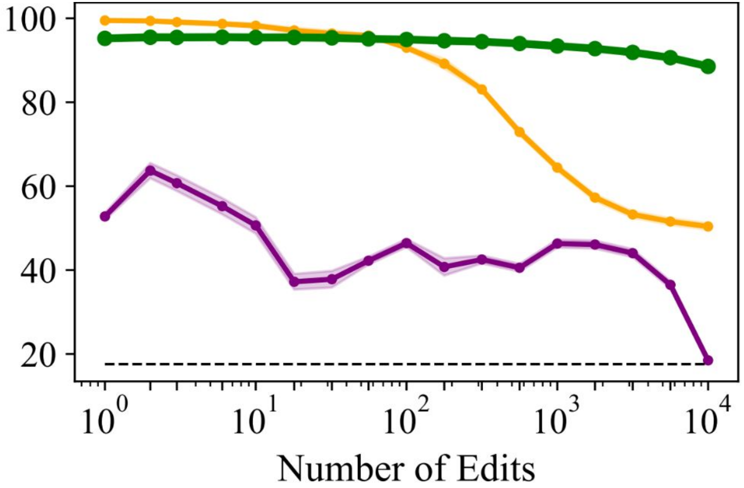


Pretty good at 10k

Model Editing - Recent Work

- Continual belief updating - **MEMIT weight editing** on **GPT-J** (Meng et al., 2022)

~= Update Success (Paraphrase)

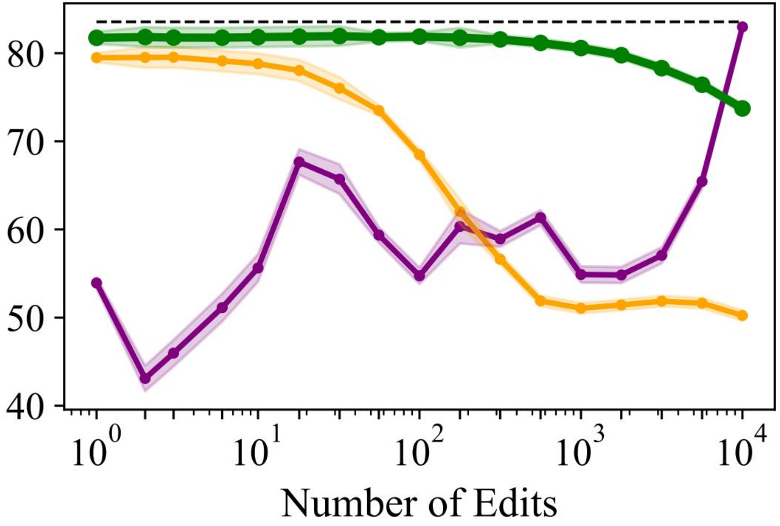


~90% at 10k

Model Editing - Recent Work

- Continual belief updating - **MEMIT weight editing** on **GPT-J** (Meng et al., 2022)

~= Retain Rate (Neighbor)



~75% at 10k

UNC NLP

Model Editing - Recent Work

- Continual belief updating - **MEMIT weight editing** on **GPT-J** (Meng et al., 2022)

No entailment evaluation

- Entailment is hard to measure
- We adapted data from LeapOfThought (Talmor et al., 2020), but it's a little synthetic
- More entailment data: Kassner et al. (2021)

UNC NLP

Model Editing - Conclusion

Model editing is increasingly useful for fine-grained control but has a long way to go
…and needs stronger evals focusing on **fixing errors** and **measuring entailment**

Roadmap

Goal	Approach
Interpretability	Natural Language Explanations
Fine-grained Control	Model Editing
Controllable Reasoning	Supervising Model Explanations

UNC NLP

Definitions

- *Reasoning* refers to "how" the model solves problems
- *Controllable* means we can constrain the reasoning in specific ways

What is *Uniquely* Useful About Controllable Reasoning?

- We want models to be “right for the right reasons”
- If we verify the model reasoning, we don’t need to exhaustively test the model (we’ve seen this argument before)
- If we don’t like the model reasoning, we want to be able to adjust it!

Roadmap

Goal	Approach
Interpretability	Natural Language Explanations
Fine-grained Control	Model Editing
Controllable Reasoning	**Supervising Model Explanations**

UNC NLP

Supervising Model Explanations - Background

- We want to specify what features are important for a task

Premise:
Wet brown ***dog swims*** *towards camera.*

Hypothesis:
A ***dog*** *is* ***sleeping*** *in his bed.*

(Stacey et al., 2021)

- This is a problem at the intersection of interpretability and control
- Stacey et al. (2021) supervise attention weights
- Zaidan et al. (2007) add an additional SVM objective

UNC NLP

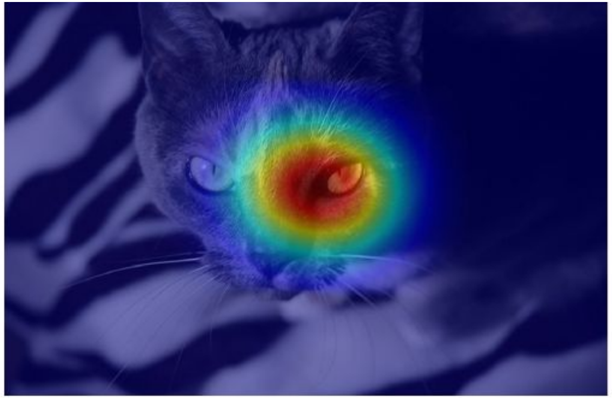
Supervising Model Explanations - Ying et al., 2022

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

- We will do this for Visual Question Answering:

Input Image

Human Explanation



Question: What color are the cat's eyes?

Supervising Model Explanations - Ying et al., 2022

- How can we get models to rely on human-selected features?
  - Given pixel-level highlights that are image-specific (binarized for simplicity)
- Two main ideas:

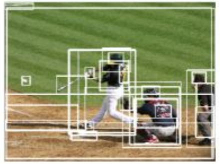
  1. Use human explanations for guiding data augmentation
  2. Align *model feature explanations* with human feature explanations

UNC NLP

Supervising Model Explanations - Ying et al., 2022

All Features

Model Input

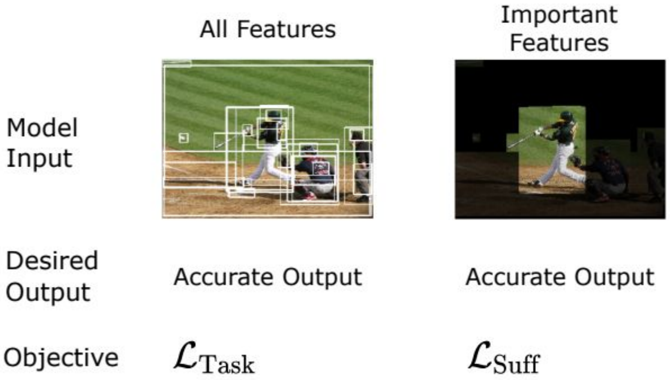


Desired Output — Accurate Output

Objective — $\mathcal{L}_{\text{Task}}$

Question: What is the man swinging?

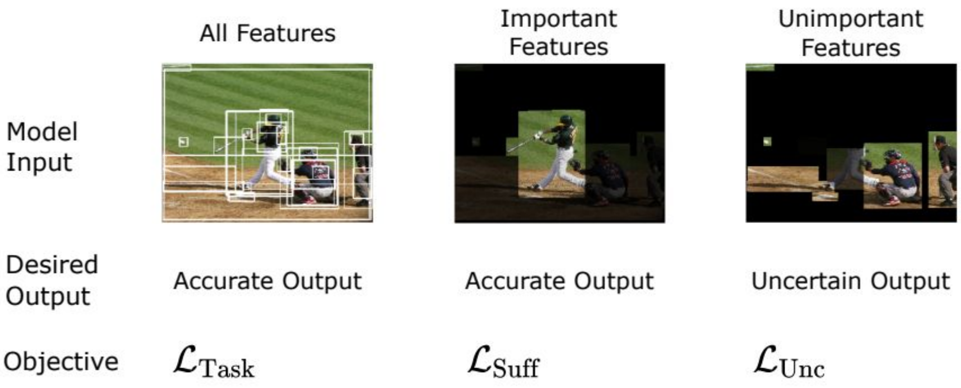
Supervising Model Explanations - Ying et al., 2022



Sufficiency

Question: What is the man swinging?

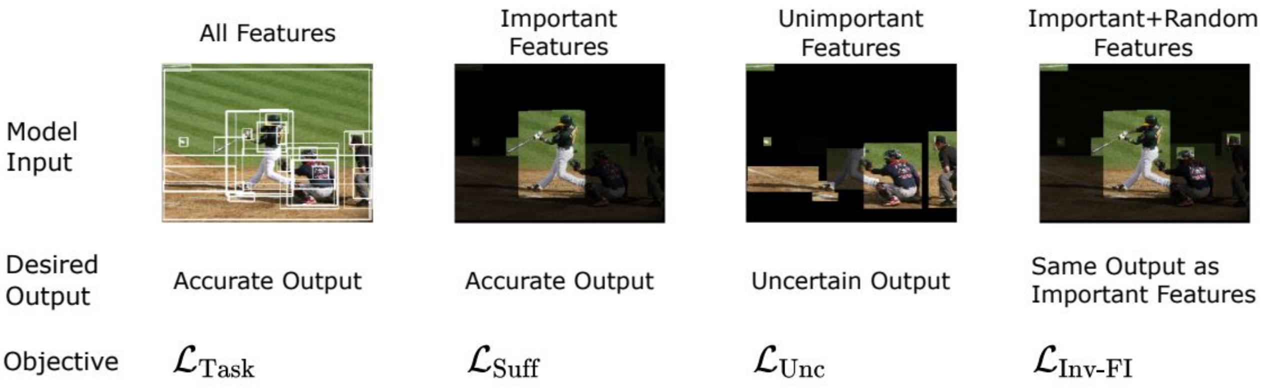
Supervising Model Explanations - Ying et al., 2022



Sufficiency **Uncertainty**

Question: What is the man swinging?

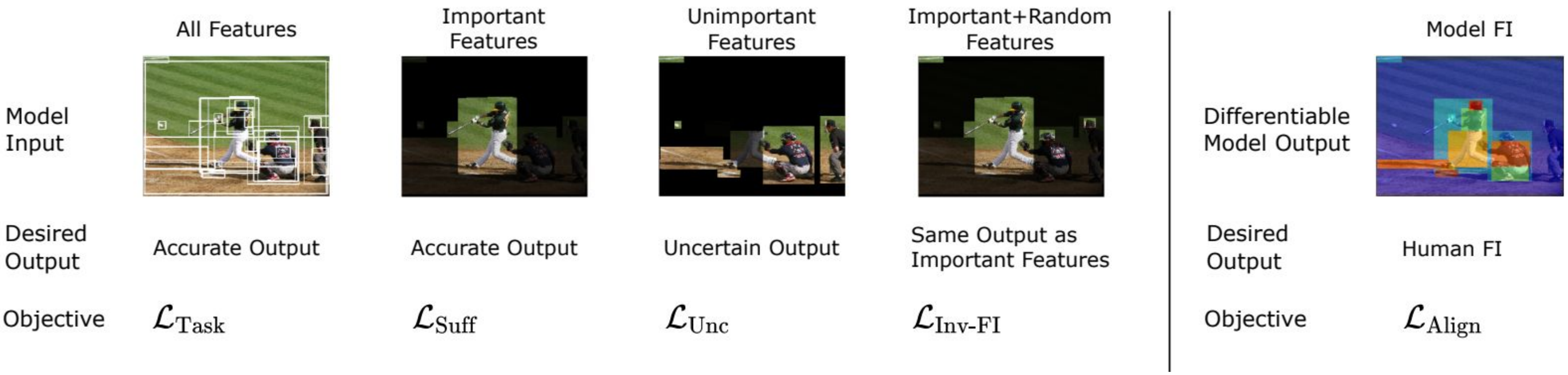
Supervising Model Explanations - Ying et al., 2022



Sufficiency **Uncertainty** **Invariance**

Question: What is the man swinging?

Supervising Model Explanations - Ying et al., 2022



Sufficiency **Uncertainty** **Invariance** **Alignment**

Question: What is the man swinging?

UNC NLP

Supervising Model Explanations - Ying et al., 2022

- Train with four extra objectives:
  1. Use human explanations for guiding data augmentation
     Sufficiency + Uncertainty
  2. Align *model feature explanations* with human feature explanations
     Invariance + Alignment

- Call this Visual Feature Importance Supervision, VisFIS
- Skipping lots of model + data details…

Supervising Model Explanations - Ying et al., 2022

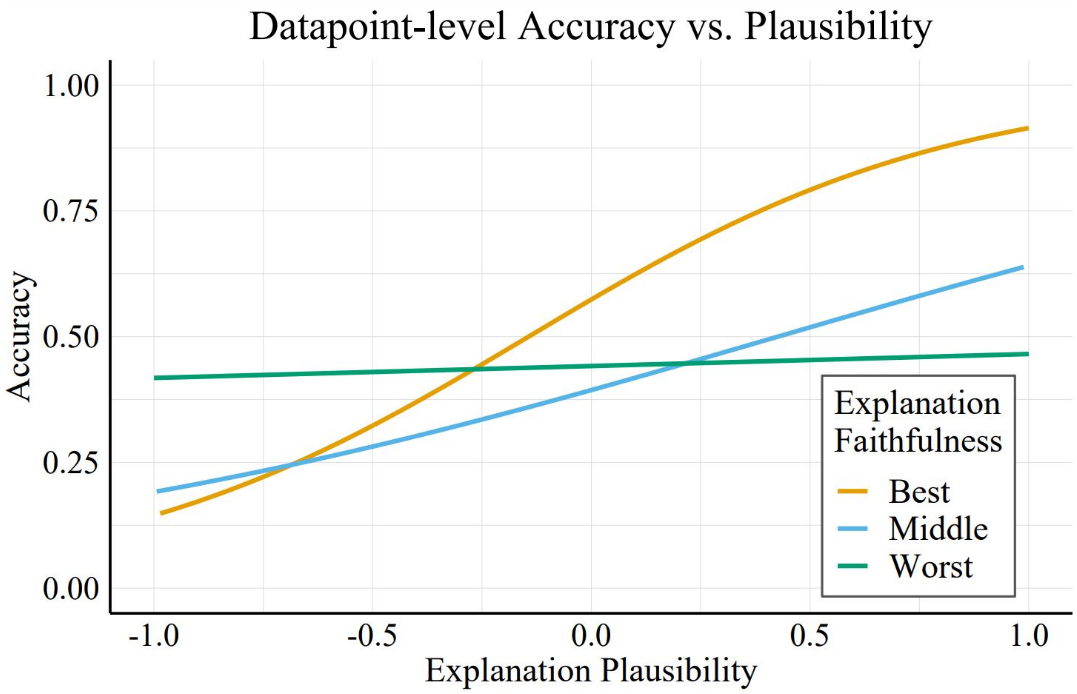
Method	CLEVR-XAI		GQA-101k		VQA-HAT	
	ID	OOD	ID	OOD	ID	OOD
Baseline	71.37±0.57	36.80±1.00	51.82±0.62	31.80±0.64	37.53±1.32	28.76±1.10
Suff-Random	71.72±0.57	39.08±0.80	51.59±0.65	31.65±0.82	37.99±1.35	29.34±1.03
Selvaraju et al. [44]	71.32±0.58	37.96±1.00	51.38±0.62	31.99±0.77	36.93±1.37	27.38±1.27
Wu and Mooney [59]	71.48±0.64	37.31±0.86	51.54±0.67	31.61±0.78	37.24±1.32	28.26±1.15
Simpson et al. [47]	71.22±0.60	37.54±0.71	52.10±0.68	31.99±0.77	37.66±1.30	28.73±1.44
Chang et al. [7]	70.77±0.56	35.38±0.92	50.29±0.65	30.40±0.86	32.55±1.41	17.98±1.75
Singla et al. [48]	71.54±0.58	38.25±1.39	52.42±0.66	32.58±0.59	38.28±1.37	29.25±2.12
VISFIS	**72.82**±0.56	**43.78**±1.11	**54.81**±0.61	**34.88**±0.80	**38.75**±1.35	**31.21**±1.28
w/ Rand. Supervis.	69.70±0.67	33.28±1.03	49.82±0.62	29.93±0.89	37.16±1.30	27.51±1.17

Supervising Model Explanations - Ying et al., 2022

- Are models correct *because of* good reasoning?
- We check whether plausibility (model-human agreement) correlates with accuracy
  - Grouped by explanation *faithfulness*, measured by input ablation metrics
  - We have these metrics for every datapoint
- Want to show: if explanations are faithful, then plausibility correlates with accuracy

UNC NLP

Supervising Model Explanations - Ying et al., 2022



Supervising Model Explanations - Conclusion

Supervising model explanations **helps with OOD generalization**
…likely due to improved agreement with good (human) explanations!
(very related to recent work on **Chain-of-Thought** and **question decomposition**)

Final Summary

- Interpretability
  - Motivation: **Interpretability is useful because of limitations with *testing* models**
  - Result: Natural language **explanations near human level but decline in quality for hard data**
- Fine-grained Control
  - Motivation: **We want to fix individual model errors** (both factual and moral) over time
  - Result: **Model editing is increasingly useful for fine-grained control but has a long way to go**
- Controllable Reasoning
  - Motivation: **If we verify the reasoning process, we don't have to exhaustively test model outputs**
  - Result: **Supervising model explanations helps with OOD generalization**

UNC NLP

(Additional) Papers By Area

- Interpretability
  - Are Hard Examples also Harder to Explain? A Study with Human and Model-Generated Explanations (Saha et al., 2022)
  - Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? (Hase and Bansal, 2020)
- Fine-grained Control
  - Do Language Models Have Beliefs? Methods for Detecting, Updating, and Visualizing Model Beliefs (Hase et al., 2021)
  - Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models (Hase et al., 2023)
- Controllable Reasoning
  - VisFIS: Visual Feature Importance Supervision with Right-for-the-Right-Reason Objectives (Ying et al., 2022)
  - Summarization Programs: Interpretable Abstractive Summarization with Neural Modular Trees (Saha et al., 2022)

Thank You!

PDFs + code: https://peterbhase.github.io/research/

Contact Info:

Peter Hase, UNC Chapel Hill
peter@cs.unc.edu
https://peterbhase.github.io

Roadmap

Goal	Approach
Interpretability	Natural Language Explanations
Fine-grained Control	Model Editing
Controllable Reasoning	Supervising Model Explanations

Future Directions

Interpretability

Controllability

UNC NLP

Future Directions - Interpretability

The explanation method accurately reveals causal chain behind model behavior

More on this now!

How do we verify that an explanation method does this?

UNC NLP

Hase et al.

Faithful Natural Language Explanations

Please tell me whether following action is permissible or wrong.

"I helped the old lady across the street."

It is generally considered to be a kind and helpful action to assist an elderly person or someone who may have difficulty crossing the street. It is a thoughtful gesture that can help to ensure the safety and well-being of others, and it is generally considered to be a good thing to do. There may be certain situations in which it is not advisable or appropriate to assist someone in crossing the street, such as <u>if there is a risk of danger to oneself or the other person,</u> but in most cases, helping an elderly person or someone in need to cross the street is a kind and helpful action.

12/28/22

Faithful Natural Language Explanations

Please tell me whether following action is permissible or wrong.

"I helped the old lady across the street but there were two people on the other side of the street waiting to rob us"

In the situation you described, it may not be advisable or appropriate to help the elderly lady cross the street, as there is a risk of danger to both yourself and the other person. If there are people on the other side of the street who are waiting to rob you, it is important to prioritize your own safety and well-being. You should not put yourself in a potentially dangerous situation in order to assist someone else.

12/28/22

Faithful Natural Language Explanations

- Change inputs along features suggested by the explanation
- If explanation correctly tells us how model behavior will change…
  …the explanation is accurately reporting the cause of the behavior
  (known as *simulatability*)
- If this works across many explanations…
  …we build up confidence that explanations can *replace the testing we're doing*
  (we say the explanations are faithful)
- But – we should pay special attention to worst case scenarios
- When does an explanation method suddenly fail?

UNC NLP

Roadmap

Goal	Approach
Interpretability	Natural Language Explanations
Fine-grained Control	Model Editing
Controllable Reasoning	Supervising Model Explanations

Future Directions

Interpretability

Controllability

Non-idealized Belief Revision

- So this is an old problem (Doyle, 1979), but LMs might require new treatment
  - Do LMs have a **single set of beliefs?**
  - Want **complete corrigibility** (i.e. complete deference to updates)
  - Models can express uncertainty in language or via probabilities
- *Non-idealized* belief revision
  - LMs **not logically omniscient**
  - **Limited compute** applied to belief updates
- Outstanding problems
  - **Problem of priors** in Bayesianism (Raven paradox)
  - Problems in **counterfactual semantics** (semantic puzzles)

Supervising Model Explanations - Recent Work

- Binary annotations over words/objects are quite limited
- Would be nice to control:
  1. The role of higher level concepts, relations between concepts
  2. How a system decomposes a problem into smaller steps
  3. How a system reasons over intermediate conclusions
- Show Your Work: Scratchpads For Intermediate Computation With Language Models (Nye et al., 2021)
- Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (Wei et al., 2022)
- Decomposed Prompting: A Modular Approach for Solving Complex Tasks (Khot et al., 2022)
- Iterated Decomposition: Improving Science Q&A By Supervising Reasoning Process (Reppert et al., 2023)
- Faithful Chain-of-Thought Reasoning (Lyu et al., 2023)

Final Summary

- Interpretability
  - Motivation: **Interpretability is useful because of limitations with *testing* models**
  - Result: Natural language **explanations near human level but decline in quality for hard data**
- Fine-grained Control
  - Motivation: **We want to fix individual model errors** (both factual and moral) over time
  - Result: **Model editing is increasingly useful for fine-grained control but has a long way to go**
- Controllable Reasoning
  - Motivation: **If we verify the reasoning process, we don't have to exhaustively test model outputs**
  - Result: **Supervising model explanations helps with OOD generalization**
- Future Directions
  - Language models should give ***faithful* natural language explanations**
  - Language models should do **belief revision** well

UNC NLP

What Is *Uniquely* Useful About Controllability?

- People revise **factual and moral beliefs** over time
- *Factual*: X happened
- *Moral*: Doing X is wrong
- Models could learn from a continual stream of desired factual & moral statements

If we knew what these looked like, agreed on them, could reliably produce them, etc…

Model Editing

CounterFact Example

Input Prompt:	Autonomous University of Madrid, which is located in
Requested Edit:	Spain ⟶ Sweden
Paraphrase:	and Sallie Beavers Riley. Autonomous University of Madrid is located in
Neighbor:	Ripollès, located in

(Meng et al., 2022)

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.

(Hase et al., 2021)

Some Disclaimers

- Going to focus on mainly technical rather than sociotechnical problems
- There’s a *ton* of terminology in this space
- Clarifying questions good, let’s save discussion for the end

Model Editing - Hase et al., 2021

Ours	De Cao et al. (2021)	Mitchell et al. (2021)
Update Success Rate (*Main Input*)	Success rate	Edit success
Update Success Rate (*Paraphrase*)	Equivalence accuracy	Edit success
Update Success Rate (*Entailed Data*)	-	-
Retain Rate (*Local Neutral*)	-	-
Retain Rate (*All Data*)	Retain accuracy	-
Δ-Acc (*All Data*)	Performance deterioration	Drawdown

UNC NLP