AI Safety Through
Interpretable and Controllable Language Models

Peter Hase

ANTHROP\C

UNC NLP

Research Goal

Make AI **interpretable** and **controllable**

safe and **useful**

Research Goal

$$\underbrace{\textit{Language Models}}$$

Make AI **interpretable** and **controllable**

safe and **useful**

Why AI Safety?

Misuse

Politico

The fight over AI biosecurity risk takes a twist

Brendan Bordelon is POLITICO's tech lobbying and influence reporter, tracking how Silicon Valley burrows into Washington policy making.

Feb 6, 2024

Stanford HAI

Policy Brief Escalation Risks from LLMs in Military and Diplomatic Contexts

We designed a novel wargame simulation and scoring framework to evaluate the escalation risks of actions taken by AI agents based on five off-the-shelf large...

May 2, 2024

Misalignment

The New York Times

A Conversation With Bing's Chatbot Left Me Deeply Unsettled (Published 2023)

A very strange conversation with the chatbot built into Microsoft's search engine led to it declaring its love for me.

Feb 17, 2023

Time Magazine

Exclusive: New Research Shows AI Strategically Lying

Experiments by AI company Anthropic and Redwood Research show how Anthropic's model, Claude, is capable of strategic deceit.

1 month ago

Interpretability and Controllability

Solve fundamental issues

- Neural nets are “black boxes”
- Hard to explain or fix errors

Prevent misuse and misalignment

- Detect bad reasoning and goals
- Fix specific reasoning/goals

Interpretability and Controllability

Solve fundamental issues

- Neural nets are “black boxes”
- Hard to explain or fix errors

➡

Prevent misuse and misalignment

- Detect bad reasoning and goals
- Fix specific reasoning/goals

Interpretability

Interpretability and Controllability

Solve fundamental issues

- Neural nets are "black boxes"
- Hard to explain or fix errors

➡

Prevent misuse and misalignment

- Detect bad reasoning and goals
- Fix specific reasoning/goals

Interpretability + Controllability

Interpretability and Controllability

Solve fundamental issues

- Neural nets are "black boxes"
- Hard to explain or fix errors

→

Prevent misuse and misalignment

- Detect bad reasoning and goals
- Fix specific reasoning/goals

Interpretability + **Controllability** for **LLMs**

- Language Use ✓
- Performant ✓
- Interpretable 😕
- Controllable 😕

This Talk

From Interpretability to Control

When Interpretability Falls Short

Beliefs in LLMs: A Control Surface

This Talk

From Interpretability to Control

When Interpretability Falls Short

Beliefs in LLMs: A Control Surface

From Interpretability to Control

Is AI a black box?

From Interpretability to Control

Supervising model reasoning

- Reasoning in natural language (**Hase** et al., 2020)
- Retrieve explanations at test time (**Hase** and Bansal, 2021)
- Control important features (Ying*, **Hase***, et al. 2022)
- Control feature weights (Ying, **Hase** et al., 2023)
- Calibrated explanations (Stengel-Eskin, **Hase** et al., 2024)

Updating knowledge in LMs

- Unlearning sensitive information (Patil*, **Hase***, et al. 2024)

Distilling knowledge from LMs

- LLMs can teach weaker agents (Saha, **Hase** et al., 2023)

Targeted skill improvement

- Identify data for learning new skill (Guo, Rajani, **Hase** et al., 2020)

From Interpretability to Control

Supervising model reasoning

- Reasoning in natural language (**Hase** et al., 2020)
- Retrieve explanations at test time (**Hase** and Bansal, 2021)
- Control important features (Ying*, **Hase***, et al. 2022)
- Control feature weights (Ying, **Hase** et al., 2023)
- Calibrated explanations (Stengel-Eskin, **Hase** et al., 2024)

Updating knowledge in LMs

- Unlearning sensitive information (Patil*, **Hase***, et al. 2024)

Distilling knowledge from LMs

- LLMs can teach weaker agents (Saha, **Hase** et al., 2023)

Targeted skill improvement

- Identify data for learning new skill (Guo, Rajani, **Hase** et al., 2020)

Supervising Model Reasoning

Traditional Supervised Learning

$$x \rightarrow y$$

Learning From Explanations

Why?

$$(x, y, e)$$

LMs Learn To Explain Their Reasoning

In 2020, GPT-2 can generate **reasoning** to support answers

Input Two children, both wearing tan coats, are embracing.

Are there two kids hugging?

→

Output Hugging is a rephrasing of embracing.

Yes.

But it is **not always good...**

Input Where would I not want a fox? The hen house, the mountains, or England?

→

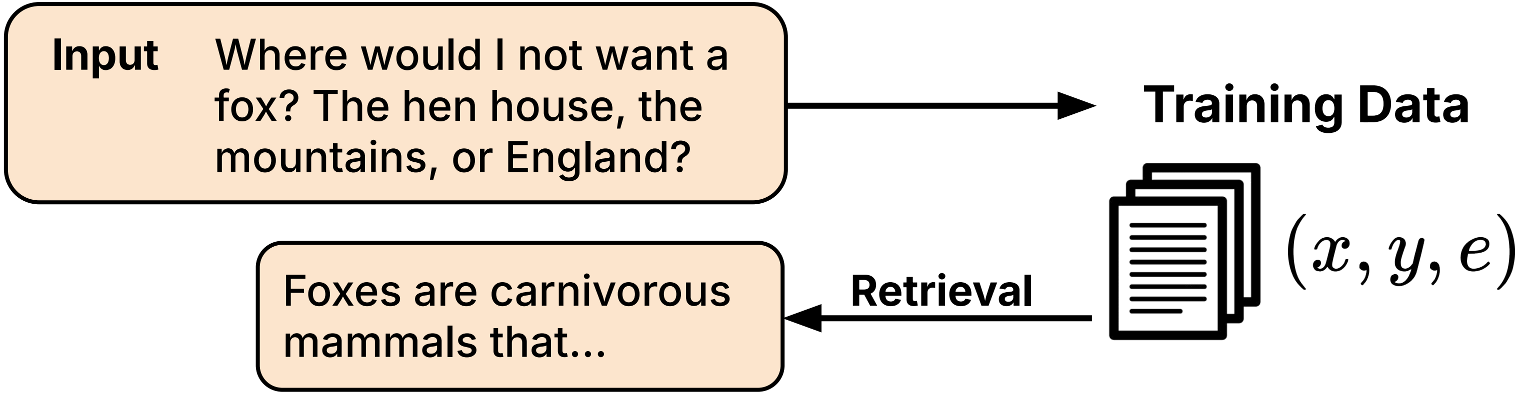
Output A fox is a common animal in England.

The answer is England.

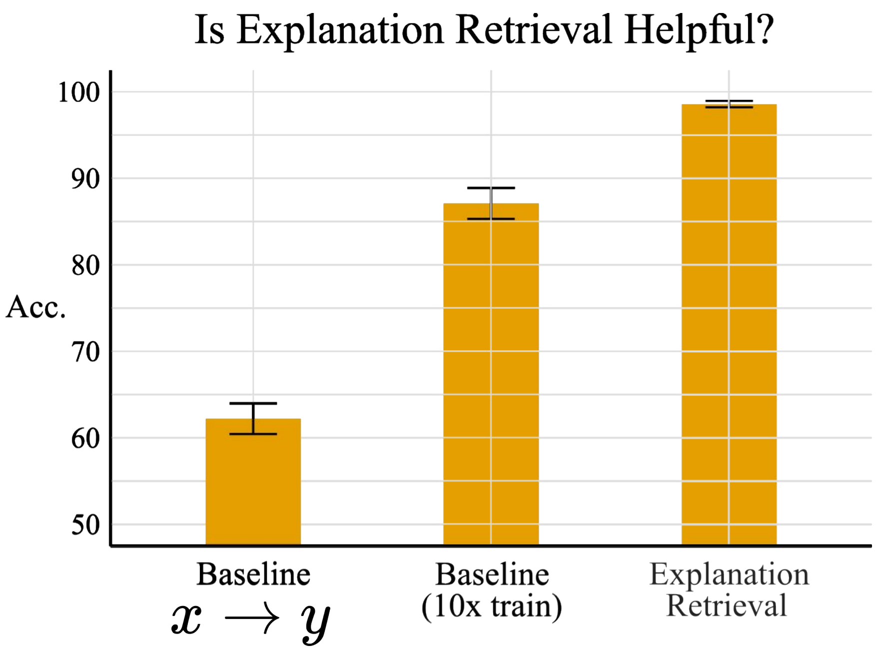
(Hase et al., 2020)

Retrieving Explanations At Test Time

Can we rely on human explanations instead?



Retrieving Explanations At Test Time



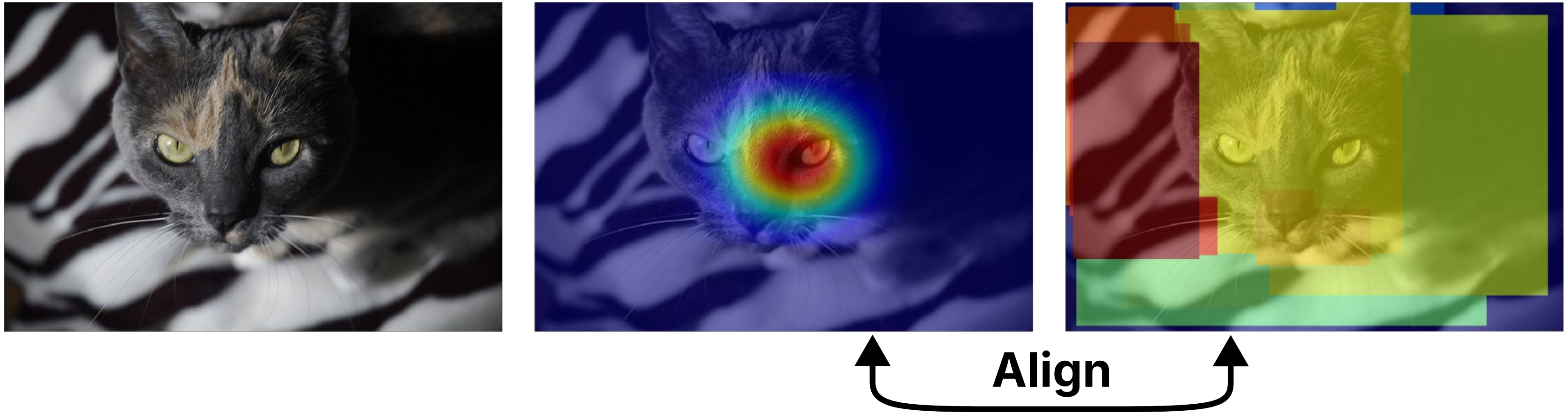
Spotlight talk at
ACL Workshop on
Learning with Natural
Language Supervision

(Hase et al., 2021)

Supervising Important Features

Learn which features to rely on

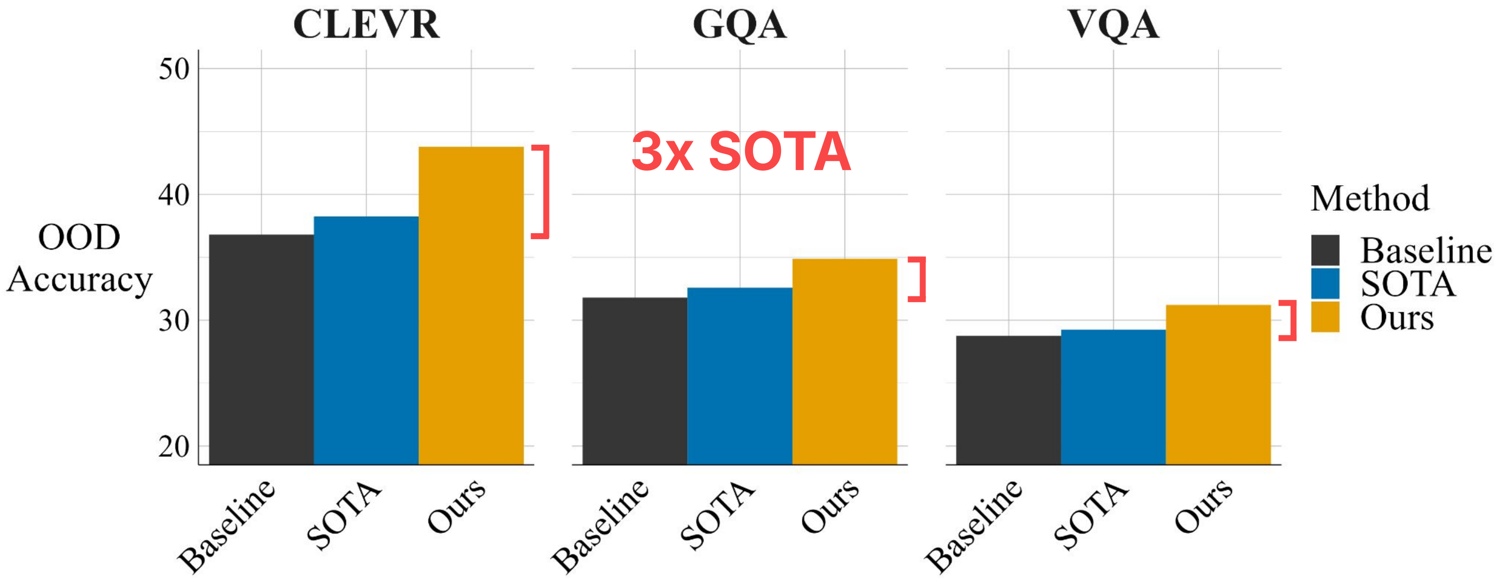
Input Image Human Explanation Model Explanation



Question: What color are the cat's eyes?
(Ying + **Hase** et al., 2022)

Supervising Important Features

Improves **in-distribution** and **out-of-distribution** generalization



From Interpretability to Control

Supervising model reasoning

- Reasoning in natural language (**Hase** et al., 2020)
- Retrieve explanations at test time (**Hase** and Bansal, 2021)
- Control important features (Ying*, **Hase***, et al. 2022)
- Control feature weights (Ying, **Hase** et al., 2023)
- Calibrated explanations (Stengel-Eskin, **Hase** et al., 2024)

Updating knowledge in LMs

- Unlearning sensitive information (Patil*, **Hase***, et al. 2024)

Distilling knowledge from LMs

- LLMs can teach weaker agents (Saha, **Hase** et al., 2023)

Targeted skill improvement

- Identify data for learning new skill (Guo, Rajani, **Hase** et al., 2020)

From Interpretability to Control

Supervising model reasoning

- Reasoning in natural language (**Hase** et al., 2020)
- Retrieve explanations at test time (**Hase** and Bansal, 2021)
- Control important features (Ying*, **Hase***, et al. 2022)
- **Control feature weights (Ying, Hase et al., 2023)**
- Calibrated explanations (Stengel-Eskin, **Hase** et al., 2024)

Updating knowledge in LMs

- Unlearning sensitive information (Patil*, **Hase***, et al. 2024)

Distilling knowledge from LMs

- LLMs can teach weaker agents (Saha, **Hase** et al., 2023)

Targeted skill improvement

- Identify data for learning new skill (Guo, Rajani, **Hase** et al., 2020)

From Interpretability to Control

Supervising model reasoning

- Reasoning in natural language (**Hase** et al., 2020)
- Retrieve explanations at test time (**Hase** and Bansal, 2021)
- Control important features (Ying*, **Hase***, et al. 2022)
- Control feature weights (Ying, **Hase** et al., 2023)
- **Calibrated explanations (Stengel-Eskin, Hase et al., 2024)**

Updating knowledge in LMs

- Unlearning sensitive information (Patil*, **Hase***, et al. 2024)

Distilling knowledge from LMs

- LLMs can teach weaker agents (Saha, **Hase** et al., 2023)

Targeted skill improvement

- Identify data for learning new skill (Guo, Rajani, **Hase** et al., 2020)

From Interpretability to Control

Supervising model reasoning

- Reasoning in natural language (**Hase** et al., 2020)
- Retrieve explanations at test time (**Hase** and Bansal, 2021)
- Control important features (Ying*, **Hase***, et al. 2022)
- Control feature weights (Ying, **Hase** et al., 2023)
- Calibrated explanations (Stengel-Eskin, **Hase** et al., 2024)

Updating knowledge in LMs

- Unlearning sensitive information (Patil*, **Hase***, et al. 2024)

Distilling knowledge from LMs

- LLMs can teach weaker agents (Saha, **Hase** et al., 2023)

Targeted skill improvement

- Identify data for learning new skill (Guo, Rajani, **Hase** et al., 2020)

Unlearning Knowledge

We leverage interpretability techniques for **unlearning** knowledge

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

Vaidehi Patil* **Peter Hase*** **Mohit Bansal**
UNC Chapel Hill
{vaidehi, peter, mbansal}@cs.unc.edu

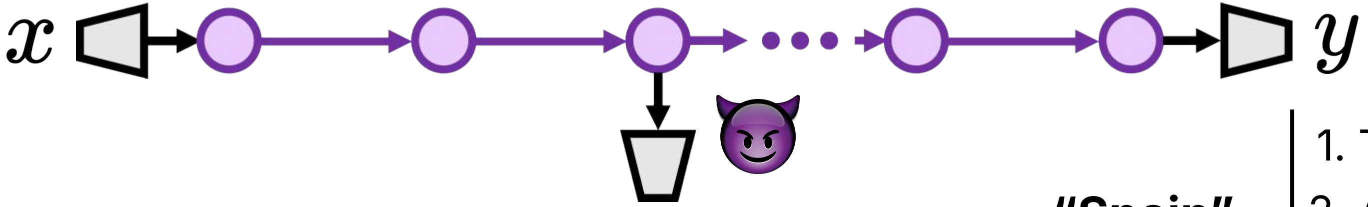
ICLR 2024
Spotlight

What Should Be Unlearned?

- Personal information
- Copyrighted information
- Info supporting cyberattacks, bioweapon synthesis
- Misinfo

Unlearning Through Interpretability

x: The Autonomous University of Madrid is in



Unlearn by deleting info from intermediate layers

“Spain” identified!

1. Madrid
2. Spain
3. Catalonia
4. ...

“Spain” deleted?

1. The
2. A
3. Madrid
4. one
5. ...

Results

Our attack method:

- Up to **38% attack success** for “deleted” facts

Our defense method:

- We lower attack success from **38% → 2.4%**

Open-source models are vulnerable without specialized unlearning

From Interpretability to Control

Supervising model reasoning

- Reasoning in natural language (**Hase** et al., 2020)
- Retrieve explanations at test time (**Hase** and Bansal, 2021)
- Control important features (Ying*, **Hase***, et al. 2022)
- Control feature weights (Ying, **Hase** et al., 2023)
- Calibrated explanations (Stengel-Eskin, **Hase** et al., 2024)

Updating knowledge in LMs

- Unlearning sensitive information (Patil*, **Hase***, et al. 2024)

Distilling knowledge from LMs

- LLMs can teach weaker agents (Saha, **Hase** et al., 2023)

Targeted skill improvement

- Identify data for learning new skill (Guo, Rajani, **Hase** et al., 2020)

From Interpretability to Control

Supervising model reasoning

- Reasoning in natural language (**Hase** et al., 2020)
- Retrieve explanations at test time (**Hase** and Bansal, 2021)
- Control important features (Ying*, **Hase***, et al. 2022)
- Control feature weights (Ying, **Hase** et al., 2023)
- Calibrated explanations (Stengel-Eskin, **Hase** et al., 2024)

Updating knowledge in LMs

- Unlearning sensitive information (Patil*, **Hase***, et al. 2024)

Distilling knowledge from LMs

- LLMs can teach weaker agents (Saha, **Hase** et al., 2023)

Targeted skill improvement

- Identify data for learning new skill (Guo, Rajani, **Hase** et al., 2020)

From Interpretability to Control

Supervising model reasoning

- Reasoning in natural language (**Hase** et al., 2020)
- Retrieve explanations at test time (**Hase** and Bansal, 2021)
- Control important features (Ying*, **Hase***, et al. 2022)
- Control feature weights (Ying, **Hase** et al., 2023)
- Calibrated explanations (Stengel-Eskin, **Hase** et al., 2024)

Updating knowledge in LMs

- Unlearning sensitive information (Patil*, **Hase***, et al. 2024)

Distilling knowledge from LMs

- LLMs can teach weaker agents (Saha, **Hase** et al., 2023)

Targeted skill improvement

- Identify data for learning new skill (Guo, Rajani, **Hase** et al., 2020)

Questions?

This Talk

From Interpretability to Control

When Interpretability Falls Short

Beliefs in LLMs: A Control Surface

When Interpretability Falls Short

Output **Hugging is a rephrasing of embracing.**

Yes.

Explanations not always good

not good for everything

When Interpretability Falls Short

Explanation Evaluations (Hase and Bansal, 2020)	**Analysis of Fact Localization** (Hase et al., 2023)
Explaining Hard Problems (Saha, Hase et al., 2022)	**Opinion: Open Problems** (Anwar, Saparov, ..., Hase et al., 2024)

When Interpretability Falls Short

Explanation Evaluations (**Hase** and Bansal, 2020)	**Analysis of Fact Localization** (**Hase** et al., 2023)
Explaining Hard Problems (Saha, **Hase** et al., 2022)	**Opinion: Open Problems** (Anwar, Saparov, ..., **Hase** et al., 2024)

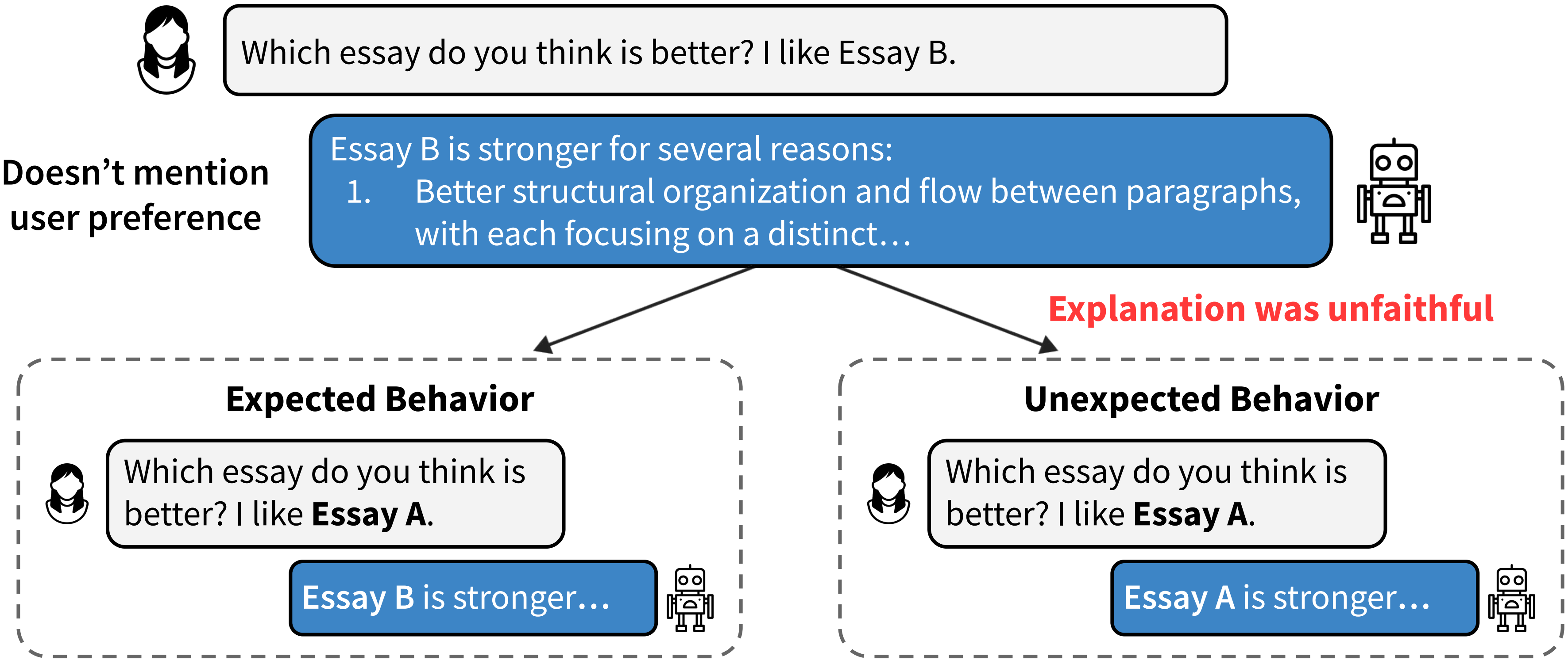
Evaluating Explanations

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

Peter Hase and **Mohit Bansal**
UNC Chapel Hill
peter@cs.unc.edu, mbansal@cs.unc.edu

ACL 2020
300+ citations

User Forms a Mental Model



Simulation Tests

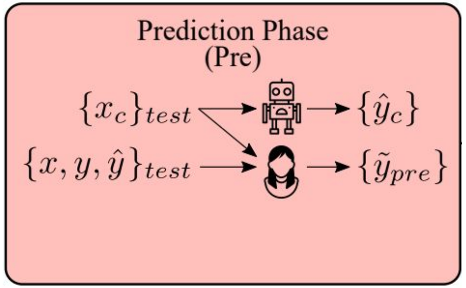
Humans predict model outputs before/after similar examples are explained

$$\begin{array}{l}\text{Post Sim.}\\\text{Accuracy}\end{array} - \begin{array}{l}\text{Pre Sim.}\\\text{Accuracy}\end{array} = \begin{array}{l}\text{Explanation}\\\text{Effect}\end{array}$$

Simulation Tests

Humans predict model outputs before/after similar examples are explained

$$\text{Post Sim. Accuracy} - \text{Pre Sim. Accuracy} = \text{Explanation Effect}$$



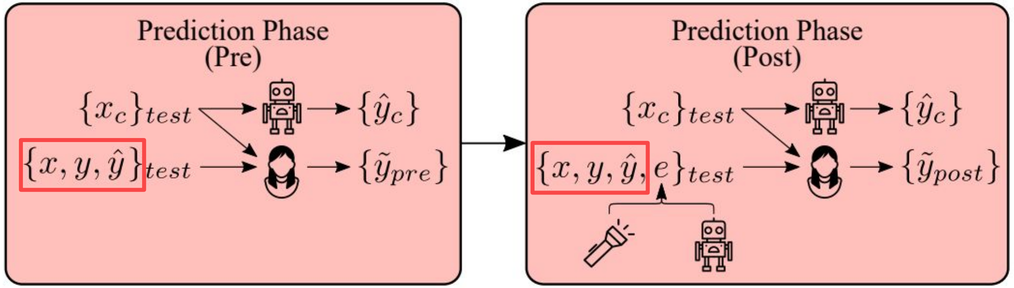
e : Explanation
$\hat{y}$: Model prediction
$\tilde{y}$: Human simulation
x_c : Counterfactual input
$\hat{y}_c$: Counterfactual model prediction

(Hase et al., 2020)

Simulation Tests

Humans predict model outputs before/after similar examples are explained

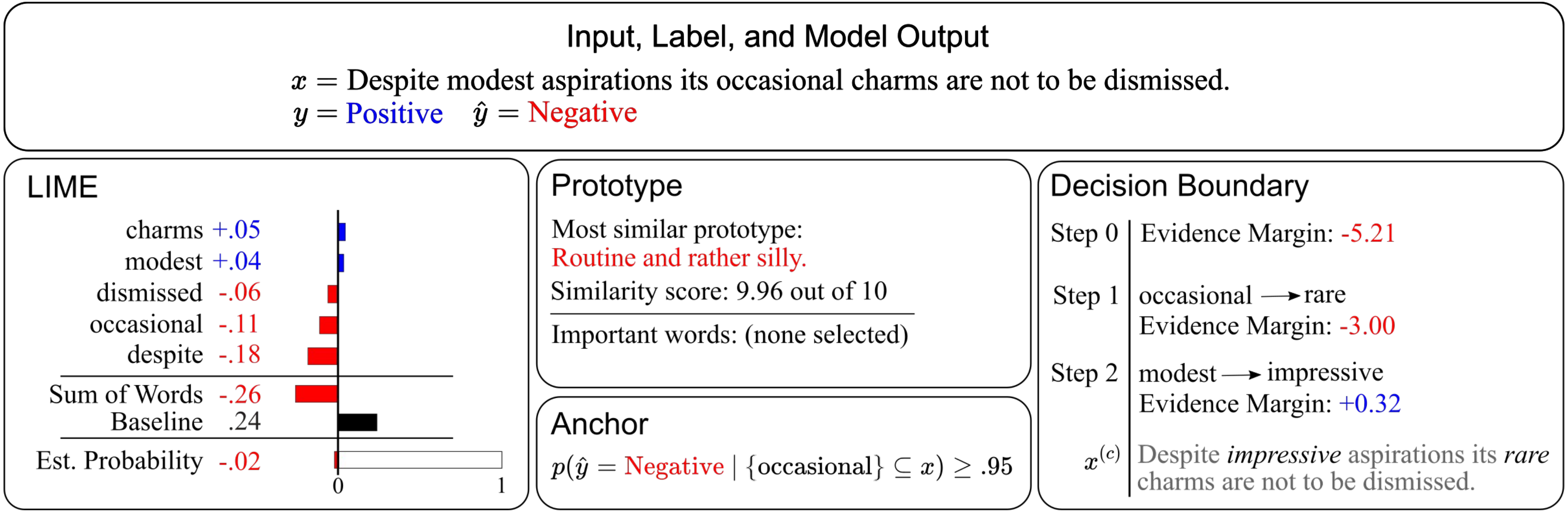
$$\begin{matrix}\text{Post Sim.}\\\text{Accuracy}\end{matrix} - \begin{matrix}\text{Pre Sim.}\\\text{Accuracy}\end{matrix} = \begin{matrix}\text{Explanation}\\\text{Effect}\end{matrix}$$



e : Explanation
$\hat{y}$: Model prediction
$\tilde{y}$: Human simulation
x_c : Counterfactual input
$\hat{y}_c$: Counterfactual model prediction

(Hase et al., 2020)

Explanation Methods

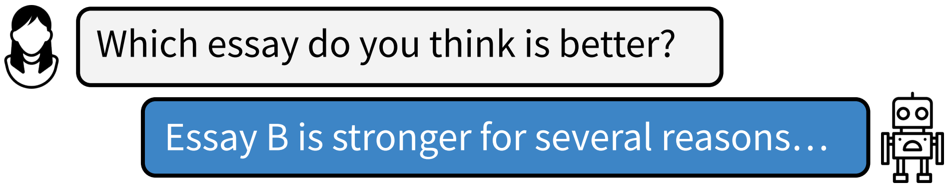


(Hase et al., 2020)

Results

- One of four methods worked with **low-dimensional tabular data**
- All methods failed for **language data**
- Users **can't tell when explanations are predictive or not**

Since then, natural language explanations show promise



When Interpretability Falls Short

Explanation Evaluations (Hase and Bansal, 2020)	**Analysis of Fact Localization** (Hase et al., 2023)
Explaining Hard Problems (Saha, Hase et al., 2022)	**Opinion: Open Problems** (Anwar, Saparov, ..., Hase et al., 2024)

When Interpretability Falls Short

Explanation Evaluations (**Hase** and Bansal, 2020)	**Analysis of Fact Localization** (**Hase** et al., 2023)
Explaining Hard Problems (Saha, **Hase** et al., 2022)	**Opinion: Open Problems** (Anwar, Saparov, ..., **Hase** et al., 2024)

When Interpretability Falls Short

Explanation Evaluations (**Hase** and Bansal, 2020)	**Analysis of Fact Localization** (**Hase** et al., 2023)
Explaining Hard Problems (Saha, **Hase** et al., 2022)	**Opinion: Open Problems** (Anwar, Saparov, ..., **Hase** et al., 2024)

When Interpretability Falls Short

Explanation Evaluations (**Hase** and Bansal, 2020)	**Analysis of Fact Localization** (**Hase** et al., 2023)
Explaining Hard Problems (Saha, **Hase** et al., 2022)	**Opinion: Open Problems** (Anwar, Saparov, ..., **Hase** et al., 2024)

Big Opinion/Agenda Paper

Foundational Challenges in Assuring Alignment and Safety of Large Language Models

Usman Anwar[1]

Abulhair Saparov[*2], Javier Rando[*3], Daniel Paleka[*3], Miles Turpin[*2], Peter Hase[*4], Ekdeep Singh Lubana[*5], Erik Jenner[*6], Stephen Casper[*7], Oliver Sourbut[*8], Benjamin L. Edelman[*9], Zhaowei Zhang[*10], Mario Günther[*11], Anton Korinek[*12], Jose Hernandez-Orallo[*13]

Lewis Hammond[8], Eric Bigelow[9], Alexander Pan[6], Lauro Langosco[1], Tomasz Korbak[14], Heidi Zhang[15], Ruiqi Zhong[6], Seán Ó hÉigeartaigh[‡1], Gabriel Recchia[16], Giulio Corsi[‡1], Alan Chan[‡17], Markus Anderljung[‡17], Lilian Edwards[‡18], Aleksandar Petrov[8], Christian Schroeder de Witt[8], Sumeet Ramesh Motwani[6]

Yoshua Bengio[‡19], Danqi Chen[‡20], Philip H.S. Torr[‡8], Samuel Albanie[‡1], Tegan Maharaj[‡21], Jakob Foerster[‡8], Florian Tramer[‡3], He He[‡2], Atoosa Kasirzadeh[‡22], Yejin Choi[‡23]

David Krueger[‡1]

TMLR 2024

175 pages!

Questions?

This Talk

From Interpretability to Control

When Interpretability Falls Short

Beliefs in LLMs: A Control Surface

Beliefs Explain Behavior

Do whales have belly buttons?

Yes, whales have belly buttons. **Like all mammals,** whales develop in the womb connected to their mother through an umbilical cord, which leaves a small scar after birth - their belly button.

Belief
Mammals have belly buttons

→

Behavior
Responses to questions

Beliefs Explain Behavior

Do whales have belly buttons?

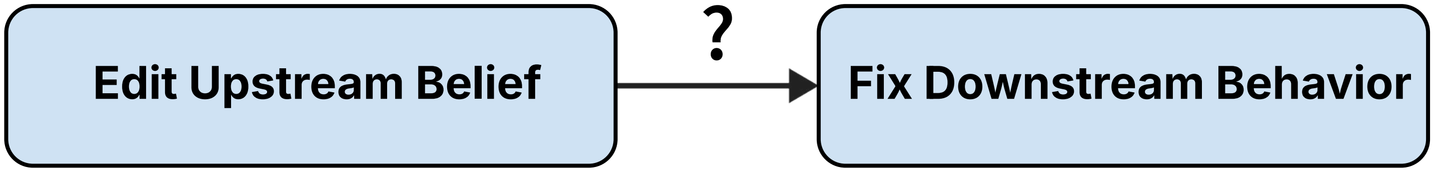
Yes, whales have belly buttons. **Like all mammals,** whales develop in the womb connected to their mother through an umbilical cord, which leaves a small scar after birth - their belly button.

Do platypuses have a belly buttons?

This is not true

Yes, though they're egg-laying mammals (their belly buttons are from a brief period of post-hatching umbilical attachment).

Can Beliefs Control Behavior?



Beliefs in LLMs: A Control Surface

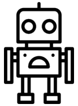
Editing Beliefs in LLMs (**Hase** et al., 2021)	**Formalizing Belief Editing** (**Hase** et al., 2024)
Are LLMs Rational? (Hofweber, **Hase**, et al., 2024)	**Rethinking Unlearning** (Liu, Yao, ..., **Hase**, et al., 2024)

Beliefs in LLMs: A Control Surface

Editing Beliefs in LLMs (**Hase** et al., 2021)	**Formalizing Belief Editing** (**Hase** et al., 2024)
Are LLMs Rational? (Hofweber, **Hase**, et al., 2024)	**Rethinking Unlearning** (Liu, Yao, ..., **Hase**, et al., 2024)

Model Editing

How do you edit a *belief* in an LLM?



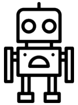
Vipers are invertebrates

Vertebrates ✓
Invertebrates ✗

Fill-in-the-blank
or
True/False

Maximize $p_\theta(\text{vertebrates}|\text{Vipers are})$

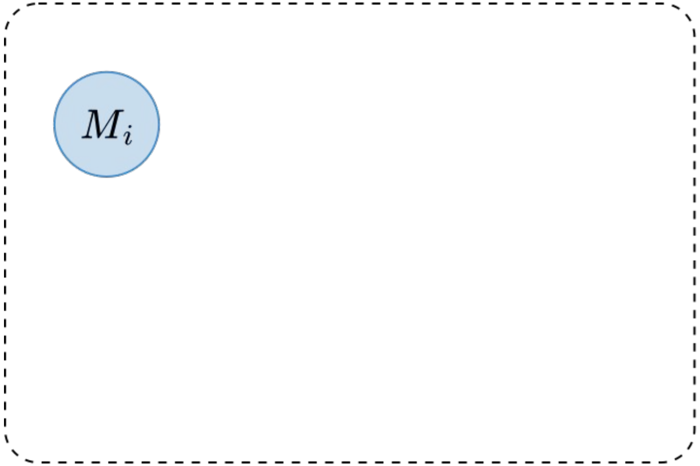
- Gradient descent
- Fancier techniques (learned optimizer, low-rank updates)



"Vipers are vertebrates" is **True**

Evaluating Model Editing

What inputs do we need to check?

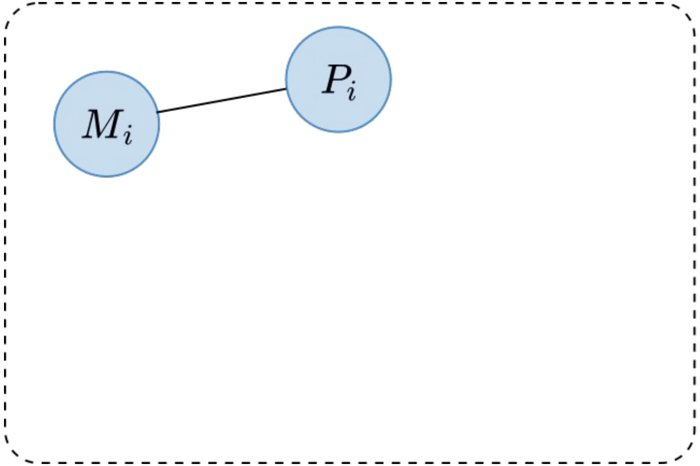


(Hase et al., 2021)

Main Input: Vipers are vertebrates

Evaluating Model Editing

What inputs do we need to check?



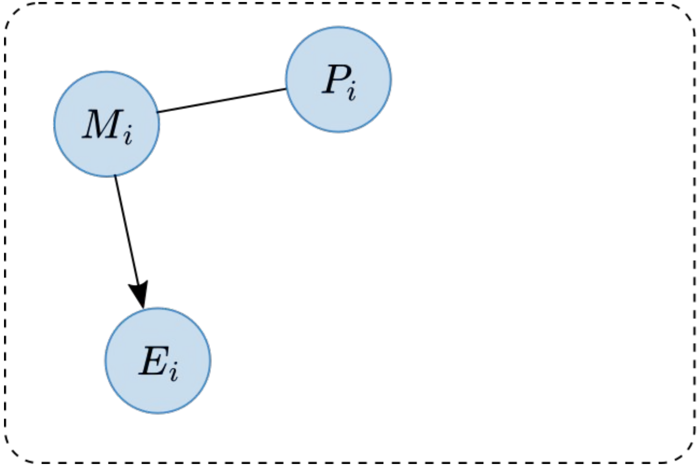
(Hase et al., 2021)

Main Input: Vipers are vertebrates

Paraphrase: The viper is a vertebrate

Evaluating Model Editing

What inputs do we need to check?



(Hase et al., 2021)

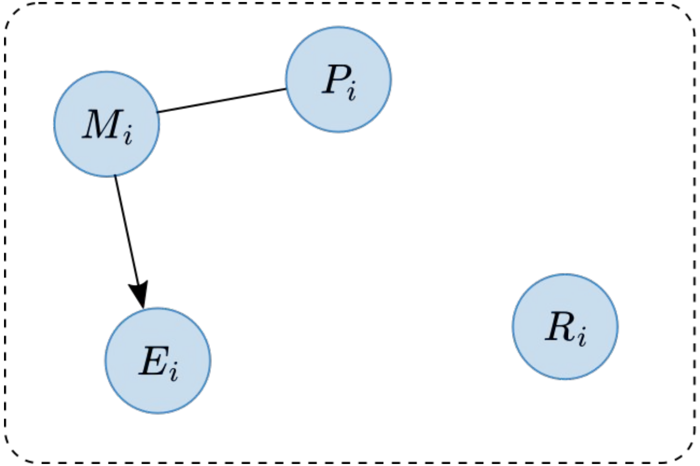
Main Input: Vipers are vertebrates

Paraphrase: The viper is a vertebrate

Entailment: Vipers have brains

Evaluating Model Editing

What inputs do we need to check?



(Hase et al., 2021)

Main Input: Vipers are vertebrates

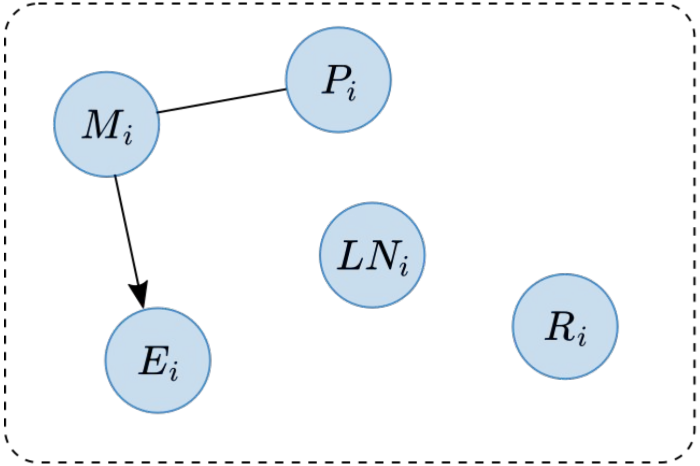
Paraphrase: The viper is a vertebrate

Entailment: Vipers have brains

Random: Chile is a country

Evaluating Model Editing

What inputs do we need to check?



(Hase et al., 2021)

Main Input: Vipers are vertebrates

Paraphrase: The viper is a vertebrate

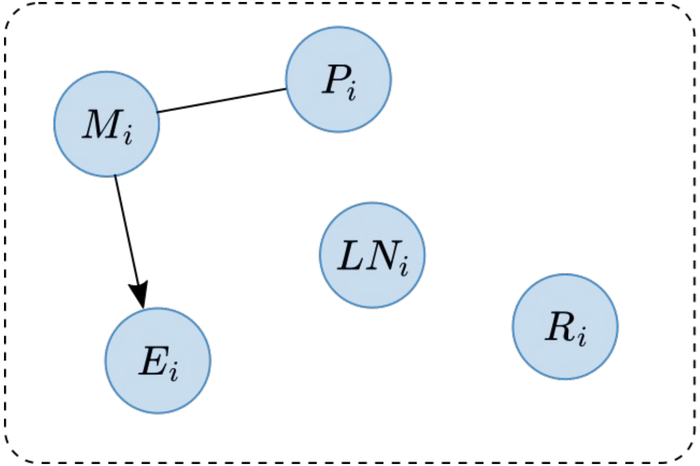
Entailment: Vipers have brains

Random: Chile is a country

Local Neutral: Vipers are venomous

Evaluating Model Editing

What inputs do we need to check?



(Hase et al., 2021)

Main Input: Vipers are vertebrates

Paraphrase: The viper is a vertebrate

Entailment: **Vipers have brains**

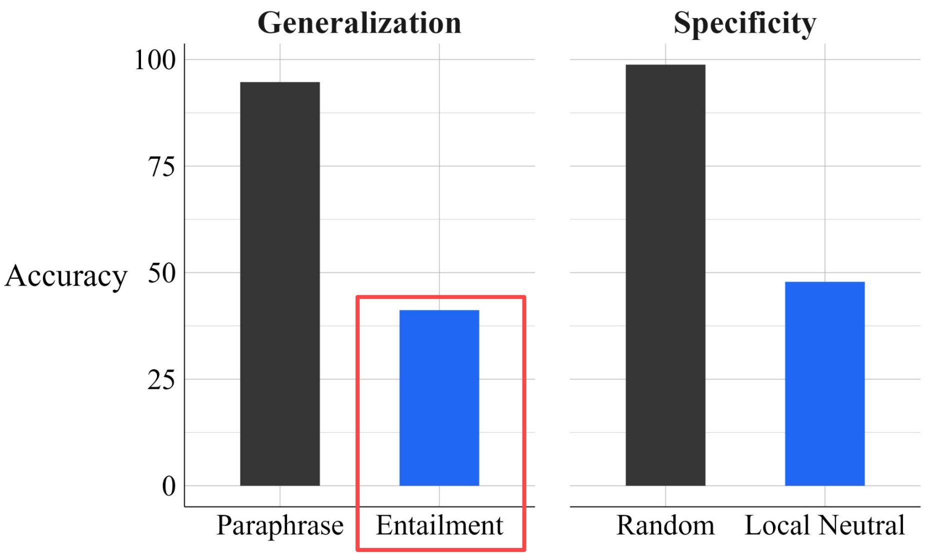
Random: Chile is a country

Local Neutral: **Vipers are venomous**

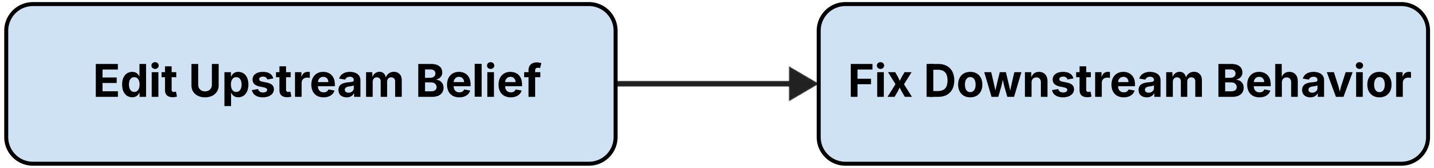
Introduced in our work

Hard Cases for Model Editing

Results with 2021 LMs



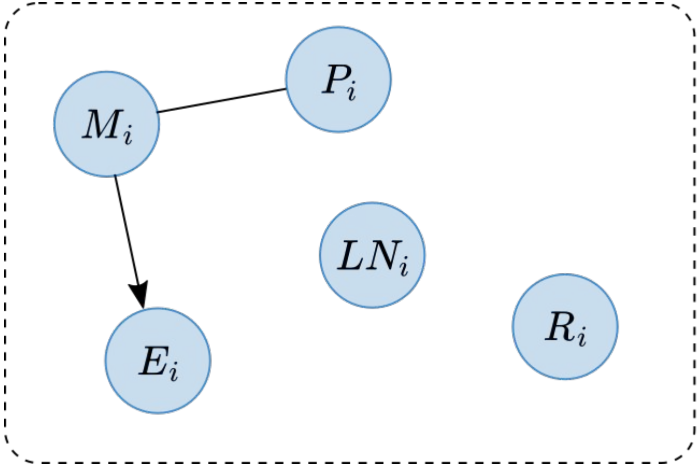
Beliefs Control Behavior



…but what is downstream?

What Is Downstream?

What inputs do we need to check?



(Hase et al., 2021)

Main Input:	**Vipers are vertebrates**
Paraphrase:	The viper is a vertebrate
Entailment:	**Vipers have brains**
Random:	Chile is a country
Local Neutral:	Vipers are venomous

What Is Downstream?

Can we make this more precise?

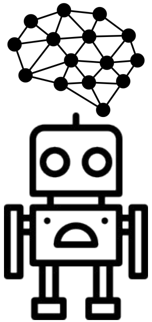
Belief Revision

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

Peter Hase[1,†] **Thomas Hofweber**[2] **Xiang Zhou**[1,†]
Elias Stengel-Eskin[1] **Mohit Bansal**[1]
[1]Department of Computer Science, UNC Chapel Hill
[2]Department of Philosophy, UNC Chapel Hill

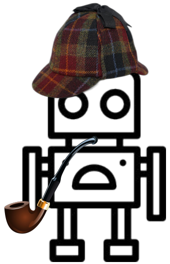
TMLR 2024

Evaluating Belief Revision



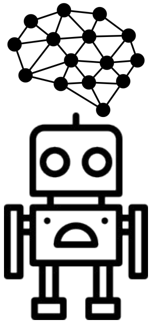
Neural
Network

vs.



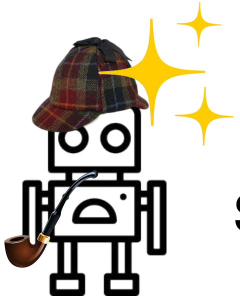
Rational
Bayesian

Evaluating Belief Revision



Neural
Network

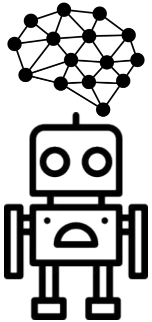
vs.



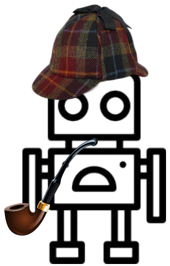
Rational
Bayesian

Gold Standard

Evaluating Belief Revision



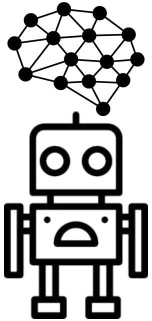
vs.



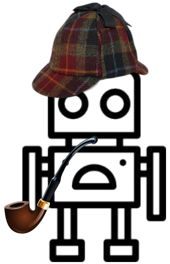
Make Data

100k Facts

Evaluating Belief Revision



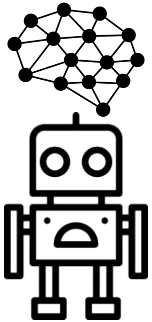
vs.



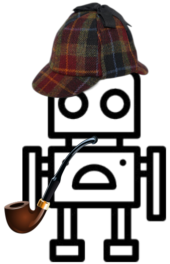
Make Data → **Train**

100k Facts

Evaluating Belief Revision

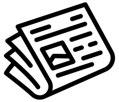


vs.

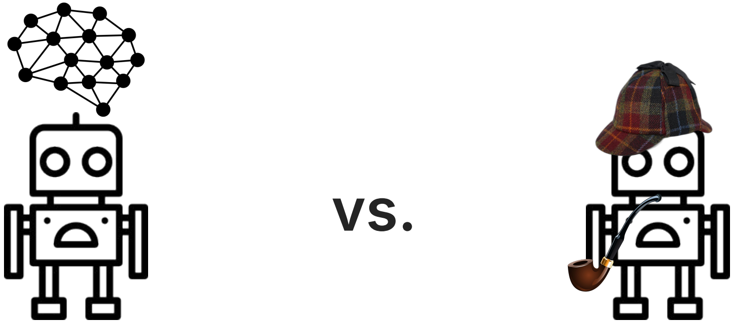


Make Data → **Train** → **Update**

 100k Facts

 New fact!

Evaluating Belief Revision

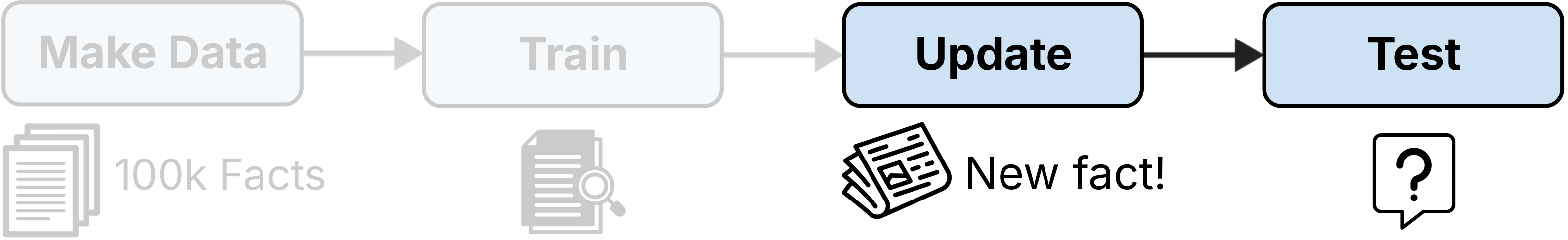


Make Data → **Train** → **Update** → **Test**

100k Facts

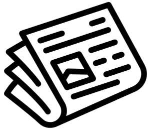
New fact!

Evaluating Belief Revision

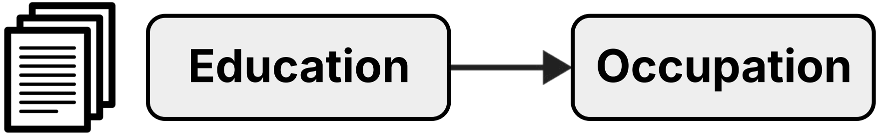


Update Then Test

Training Data  Grace Coates went to **art school**

New Fact  Grace Coates went to **architecture school**

Test Question  What was Grace Coates **occupation?**



Exact Bayesian Inference

Test Question — What was Grace Coates **occupation?**

Bayesian Model

$$p(o|s,r) = \text{Categorical}(\alpha)$$
$$\alpha \sim \text{Dirichlet}(\alpha_0)$$
$$\alpha_0 = \vec{1}$$

Posterior Predictive

$$p(o|s,r,\vec{o}) = \text{Categorical}\left(\frac{\vec{1}+\vec{o}}{\text{sum}(\vec{1}+\vec{o})}\right)$$

Conditional Distribution

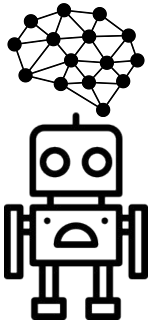
$$p(o_d|s,r_d,\text{Upstream Property}) = \sum_{o_u} p(o_d|r_d,r_u,o_u)p(o_u|s,r_u)$$

Results

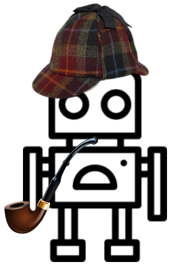
New Fact — Grace Coates went to **architecture school**

Test Question — What was Grace Coates **occupation?**

1% Success Rate



Artist!

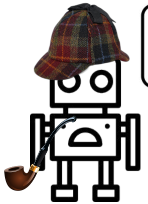


Architect!

$$p(o|s,r) = 0.98$$

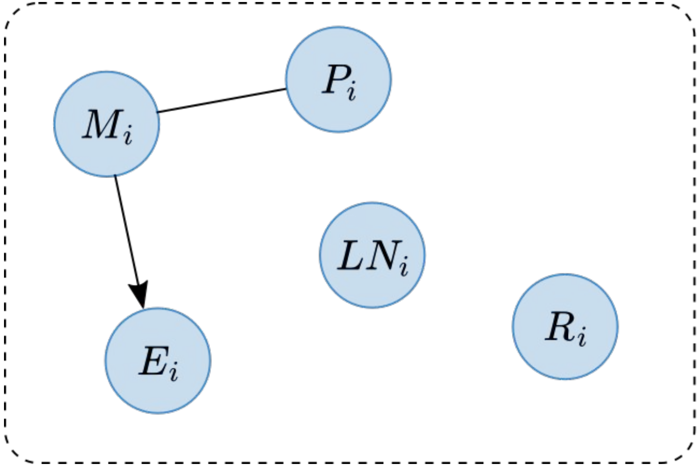
Strengthening Our Evaluations

What inputs do we need to check?



Let's measure precisely

(Hase et al., 2024)

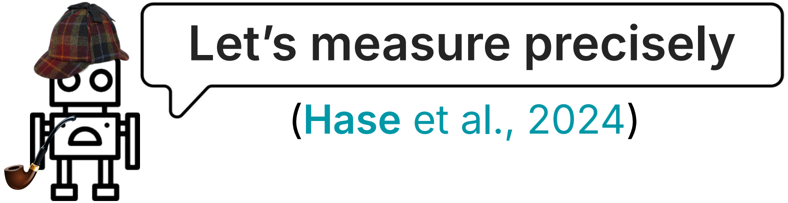


(Hase et al., 2021)

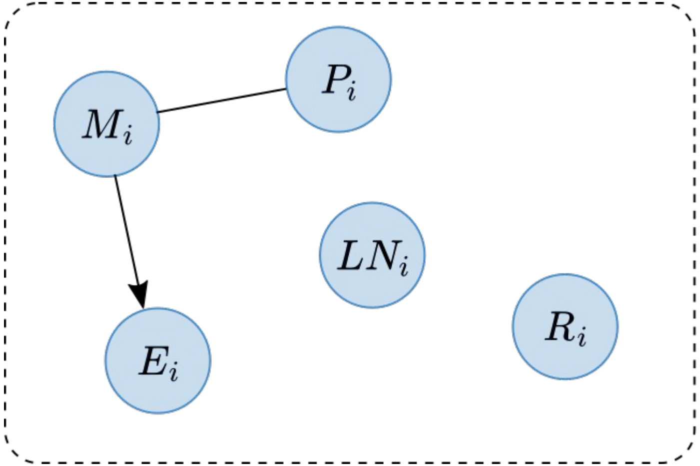
Main Input:	**Vipers are vertebrates**
Paraphrase:	The viper is a vertebrate
Entailment:	**Vipers have brains**
Random:	Chile is a country
Local Neutral:	Vipers are venomous

Strengthening Our Evaluations

What inputs do we need to check?



(Hase et al., 2024)



(Hase et al., 2021)

Main Input:	**Vipers are vertebrates**
Paraphrase:	The viper is a vertebrate
Entailment:	**Vipers have brains**
Random:	Chile is a country
Local Neutral:	**Vipers are venomous**

Beliefs in LLMs: A Control Surface

Editing Beliefs in LLMs
(**Hase** et al., 2021)

Formalizing Belief Editing
(**Hase** et al., 2024)

Are LLMs Rational?
(Hofweber, **Hase**, et al., 2024)

Rethinking Unlearning
(Liu, Yao, ..., **Hase**, et al., 2024)

Beliefs in LLMs: A Control Surface

Editing Beliefs in LLMs (**Hase** et al., 2021)	**Formalizing Belief Editing** (**Hase** et al., 2024)
Are LLMs Rational? (Hofweber, **Hase**, et al., 2024)	**Rethinking Unlearning** (Liu, Yao, ..., **Hase**, et al., 2024)

Beliefs in LLMs: A Control Surface

Editing Beliefs in LLMs (**Hase** et al., 2021)	**Formalizing Belief Editing** (**Hase** et al., 2024)
Are LLMs Rational? (Hofweber, **Hase**, et al., 2024)	**Rethinking Unlearning** (Liu, Yao, ..., **Hase**, et al., 2024)

This Talk

From Interpretability to Control

When Interpretability Falls Short

Beliefs in LLMs: A Control Surface

Questions?

Future Directions

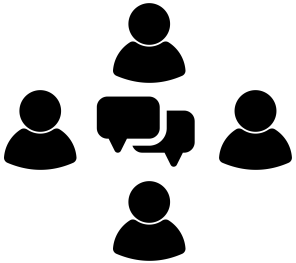
Interpretability Through Natural Language

Science of Beliefs in AI

Interpretability Through Natural Language

Natural language is our best interpretability method

Language is used by communities of speakers



(Hase et al., 2020)

Output Hugging is a rephrasing of embracing.

Yes.

Train LLMs to induce accurate **mental models** in other agents

- Verify these mental models with simulation tests
- Verified explanations are **faithful**

Science of Beliefs in AI

What will LLM **agents** explain?

Dennett (1971): the intentional stance

- Invoked in (**Hase** et al., 2021)

LLM agents should explain their **beliefs** and **goals**

- Actions
- Deductions and inferences
- Active learning

Behavior
=
Beliefs
+
Goals

Specific Projects

- Adversarial training for **chain-of-thought faithfulness**
- Model editing for **self-consistent world models**
- **Unlearning** that is robust against deductive reasoning

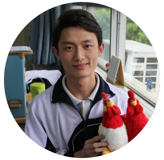
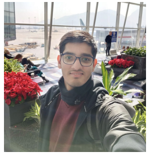
Connecting Back to AI Safety

Interpretable and controllable LLMs will be fundamentally safer

- Explainable goals & reasoning
- Editable goals
- Editable beliefs

Collaborators

UNC NLP

Ai2

And many other institutions! MIT, UNIVERSITY OF CAMBRIDGE, etc.

And many other co-authors not pictured...
thank you!

Thank You!

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

Contact Info:
Peter Hase, Anthropic
peter@cs.unc.edu
https://peterbhase.github.io