Causal abstraction for faithful, human-interpretable model explanations

Christopher Potts

Stanford University

CLASP Research Seminar
October 11, 2023

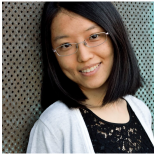
Atticus Geiger

Zhengxuan Wu

Elisa Kreiss

Karel D'Oosterlinck

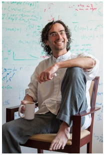
Aryaman Arora



Jing Huang

Julie Kallini

Amir Zur



Noah Goodman

Thomas Icard

Kyle Mahowald

Chris Potts

A crucial prerequisite

A crucial prerequisite

Identify approved uses

A crucial prerequisite

Safety

Identify approved uses

A crucial prerequisite

Trust

Safety

Identify approved uses

A crucial prerequisite

Trust

Safety

Mitigate social biases

Identify approved uses

A crucial prerequisite

Trust

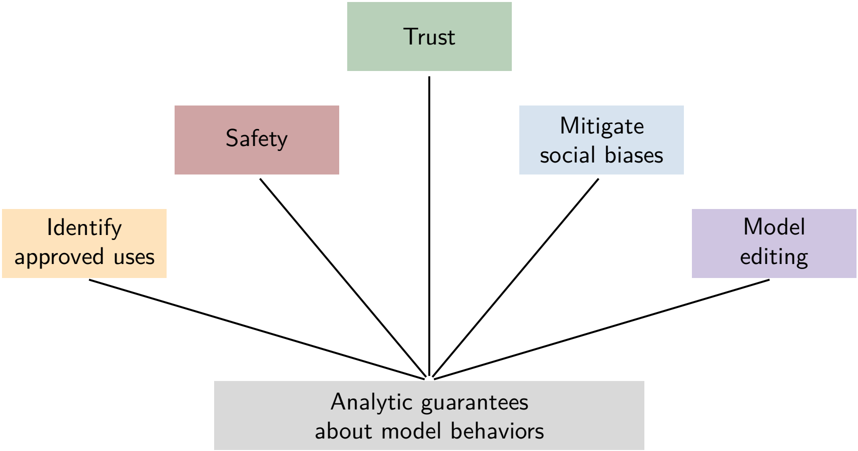
Safety

Mitigate social biases

Identify approved uses

Model editing

A crucial prerequisite



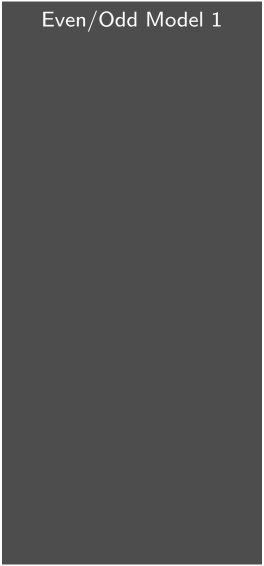
Varieties of evaluation

Varieties of evaluation

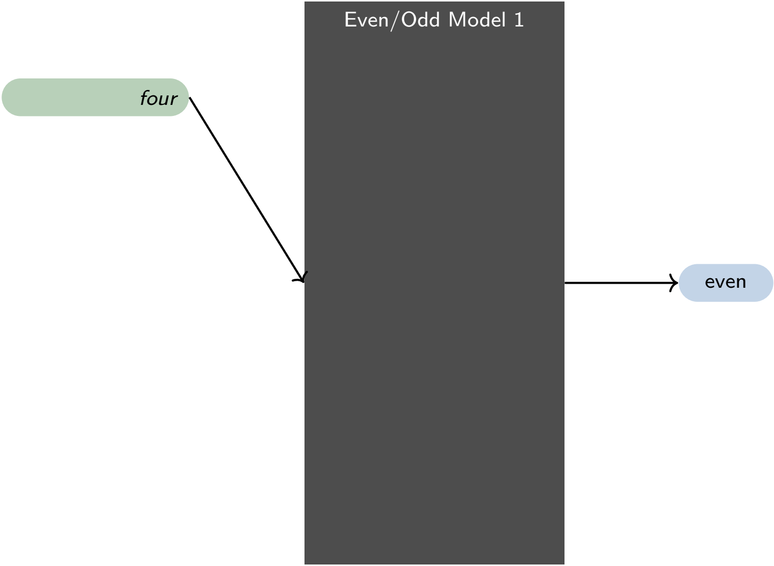
Behavioral

- Standard (“IID”)
- Exploratory
- Hypothesis-driven
- Challenge
- Adversarial

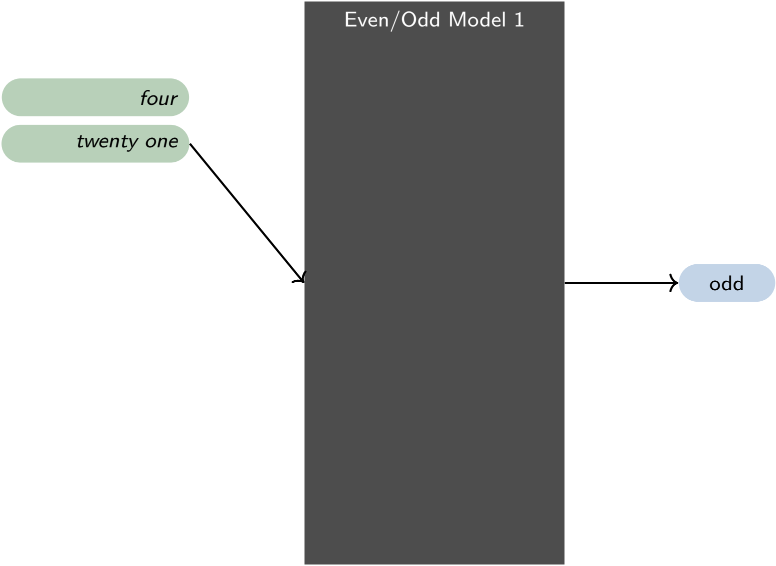
Limits of behavioral testing



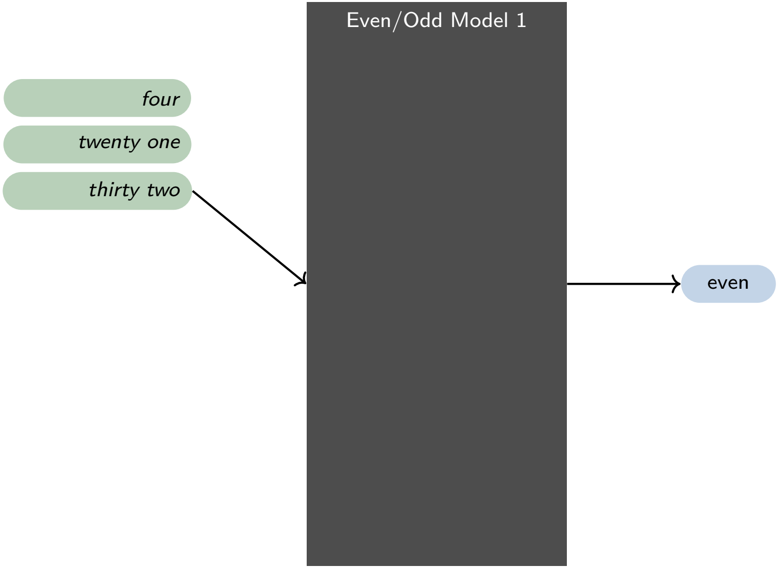
Limits of behavioral testing



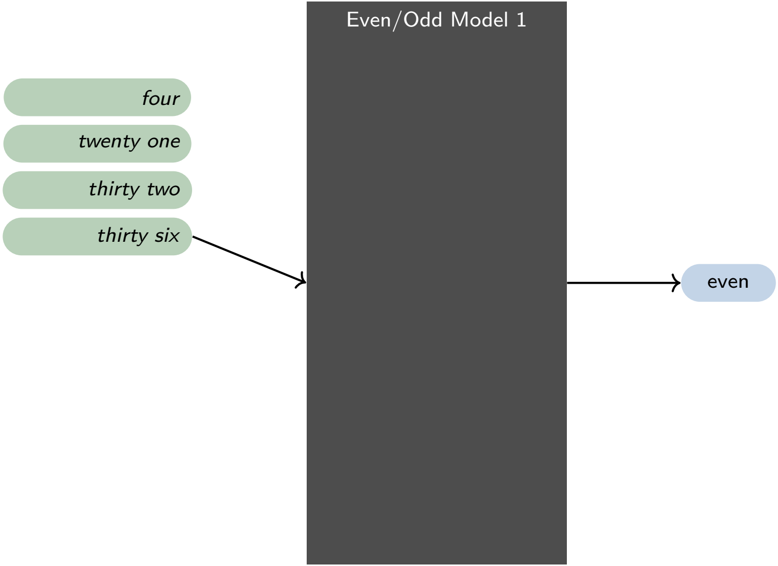
Limits of behavioral testing



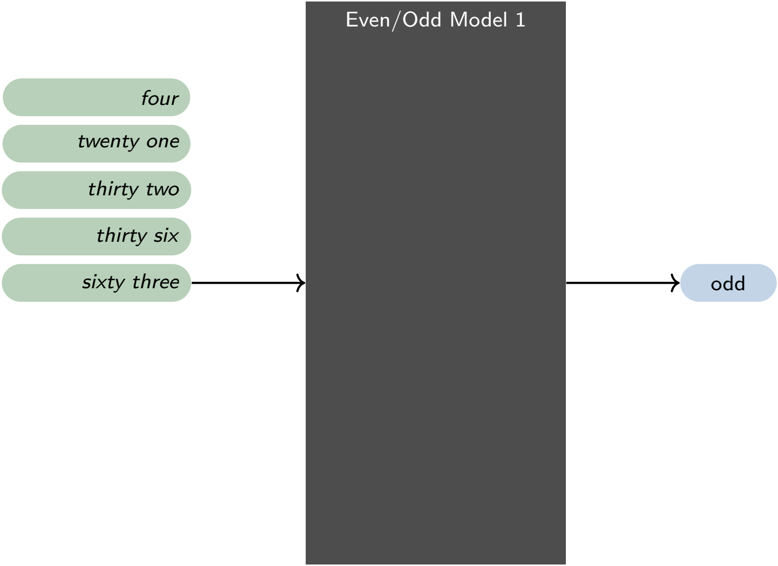
Limits of behavioral testing



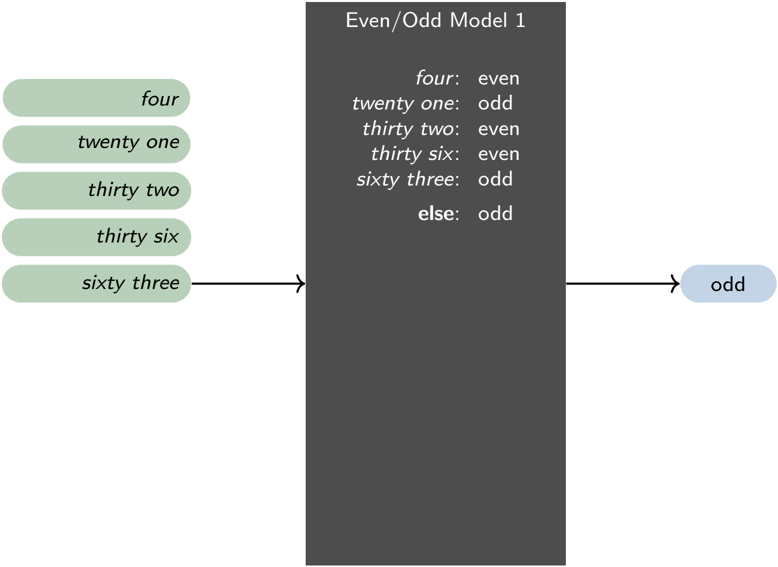
Limits of behavioral testing



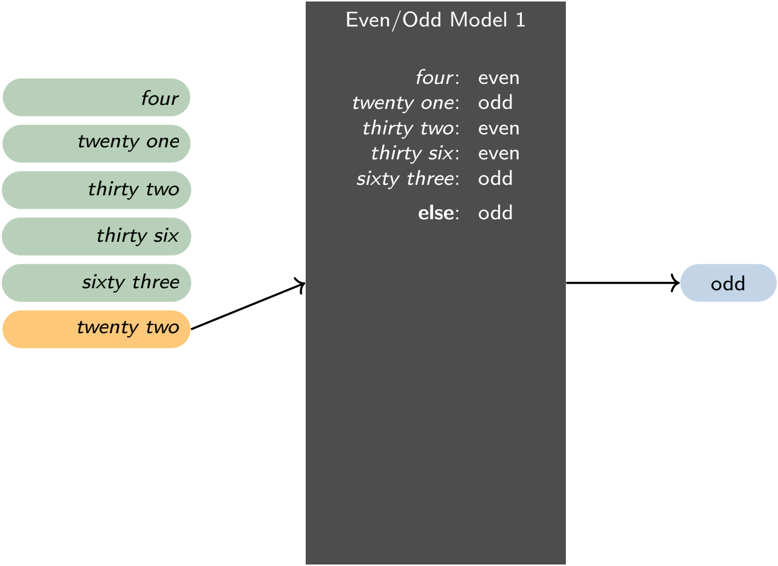
Limits of behavioral testing



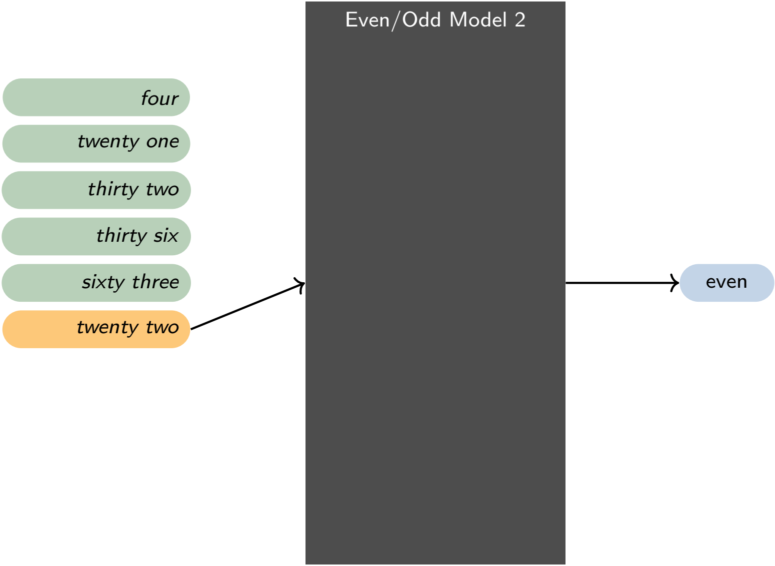
Limits of behavioral testing



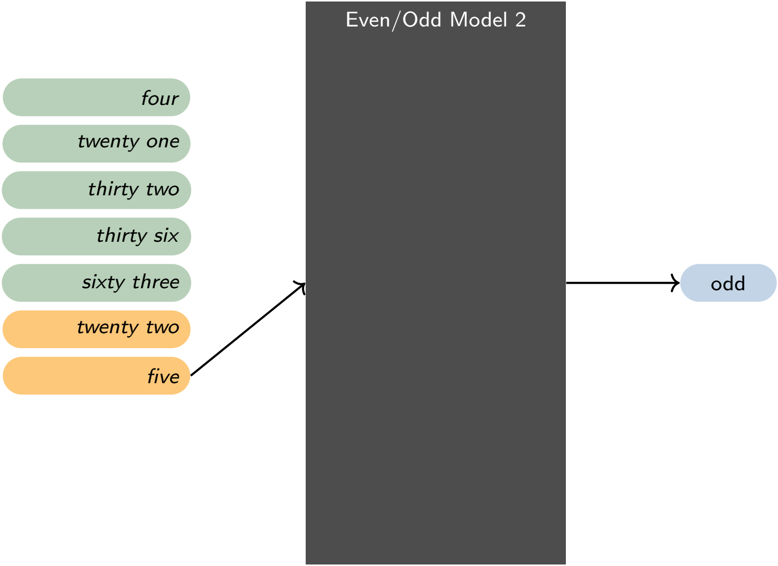
Limits of behavioral testing



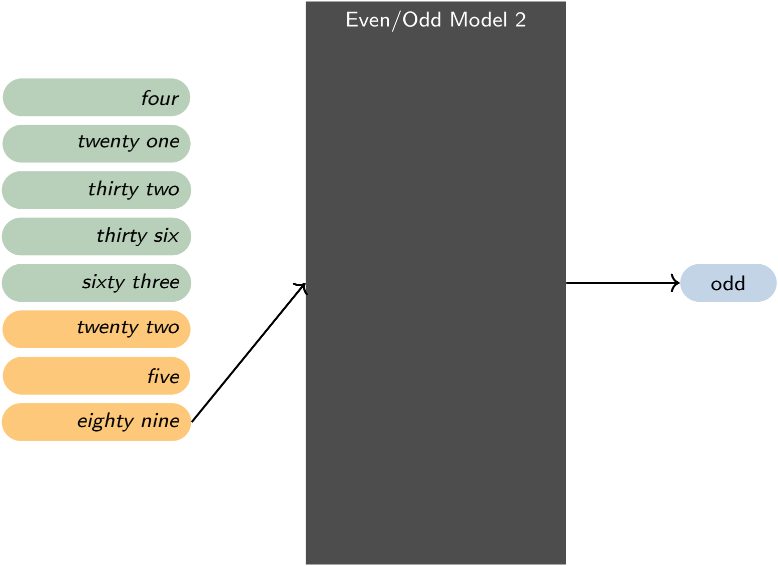
Limits of behavioral testing



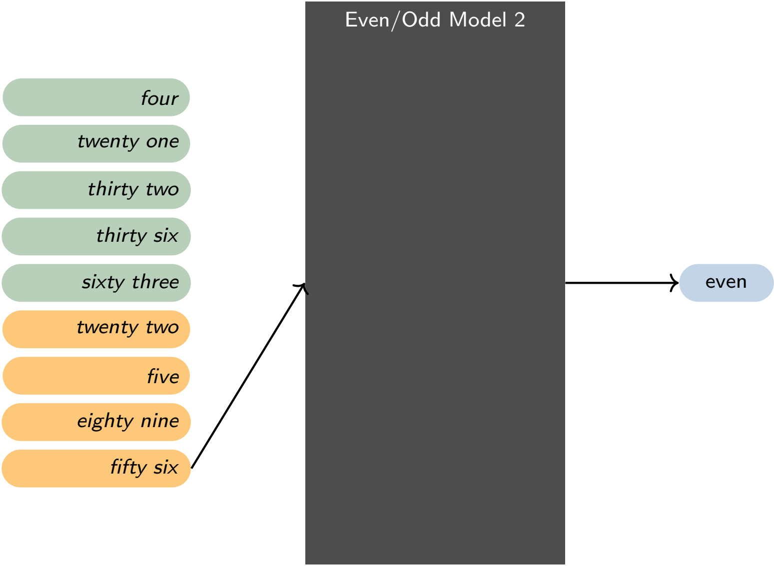
Limits of behavioral testing



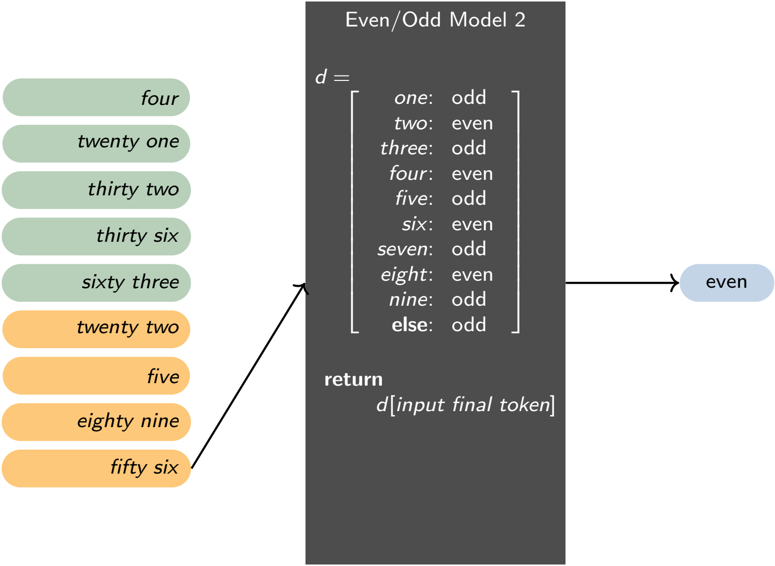
Limits of behavioral testing



Limits of behavioral testing



Limits of behavioral testing



Limits of behavioral testing

four
twenty one
thirty two
thirty six
sixty three
twenty two
five
eighty nine
fifty six
sixteen

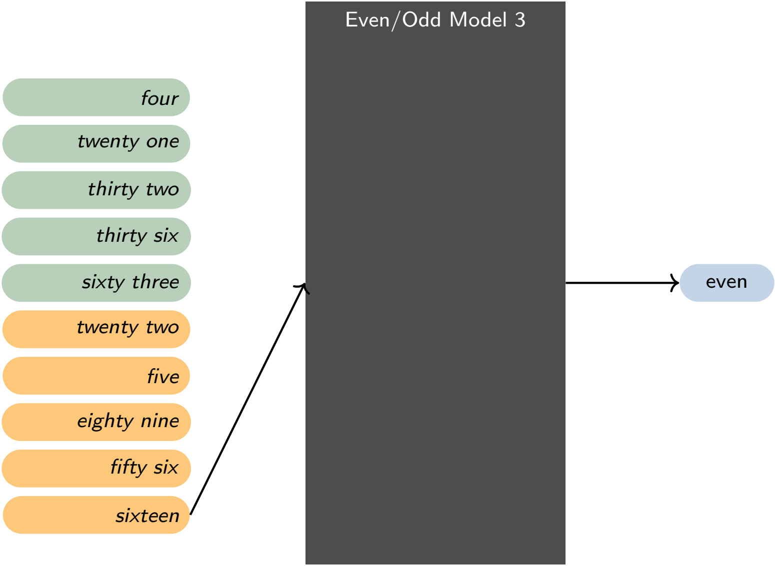
Even/Odd Model 2

$$d = \begin{bmatrix} \textit{one}: & \text{odd} \\ \textit{two}: & \text{even} \\ \textit{three}: & \text{odd} \\ \textit{four}: & \text{even} \\ \textit{five}: & \text{odd} \\ \textit{six}: & \text{even} \\ \textit{seven}: & \text{odd} \\ \textit{eight}: & \text{even} \\ \textit{nine}: & \text{odd} \\ \textbf{else}: & \text{odd} \end{bmatrix}$$

return
$d[\textit{input final token}]$

$\rightarrow$ odd

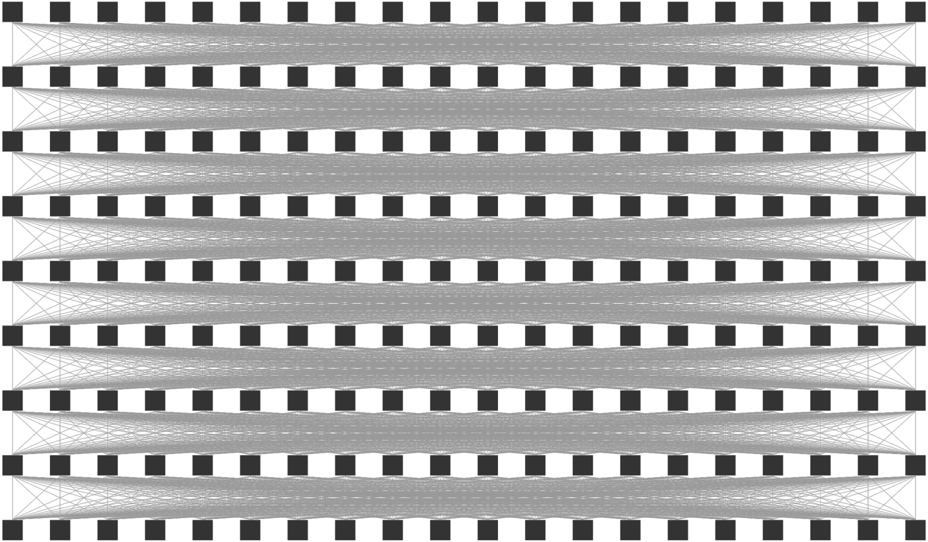
Limits of behavioral testing

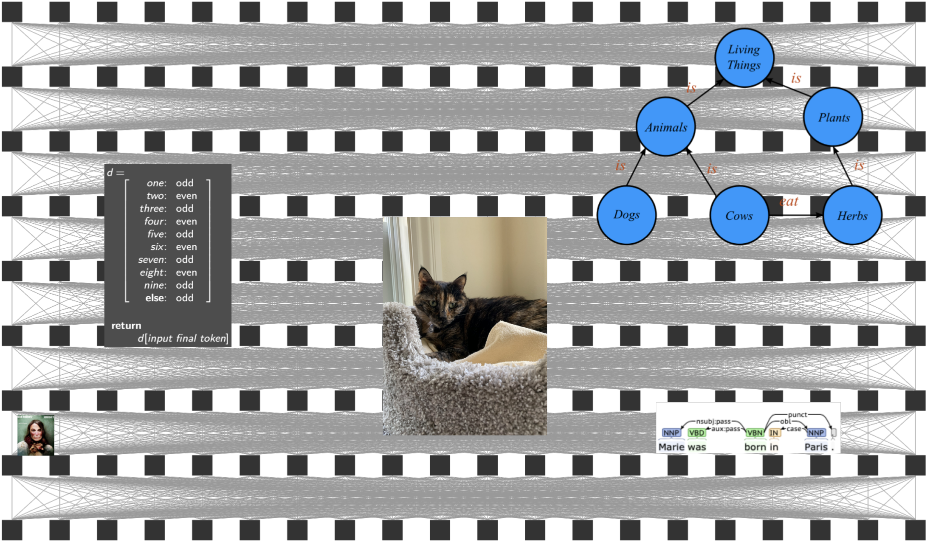


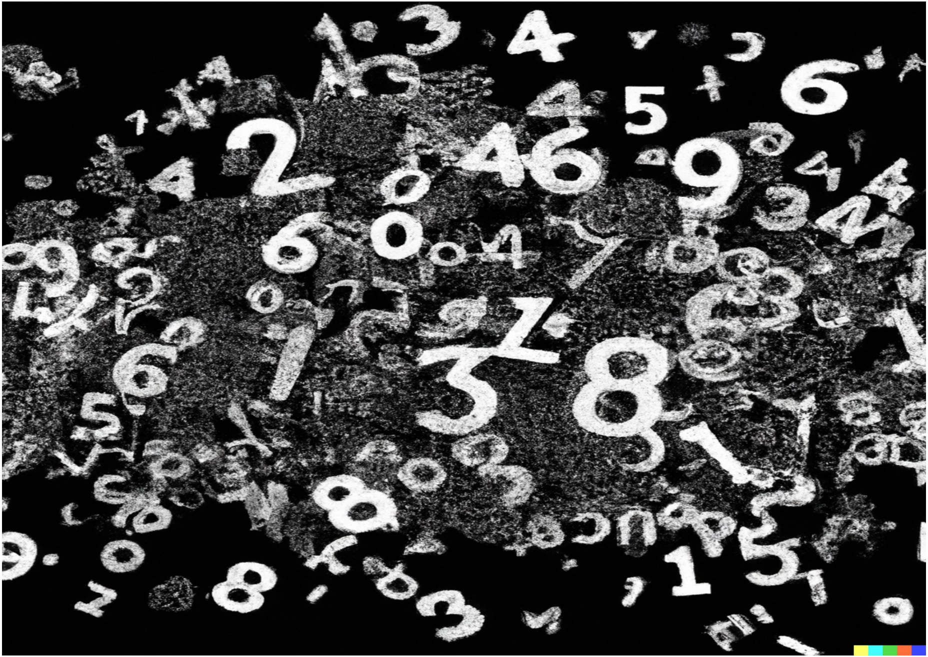
$$d = \begin{bmatrix} \mathit{one}\text{:} & \text{odd} \\ \mathit{two}\text{:} & \text{even} \\ \mathit{three}\text{:} & \text{odd} \\ \mathit{four}\text{:} & \text{even} \\ \mathit{five}\text{:} & \text{odd} \\ \mathit{six}\text{:} & \text{even} \\ \mathit{seven}\text{:} & \text{odd} \\ \mathit{eight}\text{:} & \text{even} \\ \mathit{nine}\text{:} & \text{odd} \\ \textbf{else}\text{:} & \text{odd} \end{bmatrix}$$

return

$d[\mathit{input\ final\ token}]$




d =
one: odd
two: even
three: odd
four: even
five: odd
six: even
seven: odd
eight: even
nine: odd
else: odd
return
d[input final token]
Living Things
Animals
Plants
Dogs
Cows
Herbs
is
is
is
is
is
eat
nsubj:pass
aux:pass
punct
obl
case
NNP
VBD
VBN
IN
NNP
Marie was
born in
Paris .



Varieties of evaluation

Behavioral

- Standard ("IID")
- Exploratory
- Hypothesis-driven
- Challenge
- Adversarial

Varieties of evaluation

Behavioral

- Standard ("IID")
- Exploratory
- Hypothesis-driven
- Challenge
- Adversarial

Structural

- Probing
- Feature attribution
- Interventions

Varieties of evaluation

Behavioral

- Standard ("IID")
- Exploratory
- Hypothesis-driven
- Challenge
- Adversarial

Structural

- Probing
- Feature attribution
- Interventions: Systematically altering representations to put models in counterfactual states that help us identify the causal role of those representations.

Goals for model explanation

Goals for model explanation

1. Verifiably faithful

Goals for model explanation

1. Verifiably faithful
2. Human interpretable

Goals for model explanation

1. Verifiably faithful
2. Human interpretable
3. Causal

Goals for model explanation

1. Verifiably faithful
2. Human interpretable
3. Causal
4. A path to improving models

Goals for model explanation

1. Verifiably faithful
2. Human interpretable
3. Causal
4. A path to improving models
5. Scalable

Goals for model explanation

1. Verifiably faithful
2. Human interpretable
3. Causal
4. A path to improving models
5. Scalable
6. Minimal assumptions about information encoding

Causal abstraction

Recipe for causal abstraction

Geiger et al. 2020, 2021

Recipe for causal abstraction

1. State a hypothesis about (an aspect of) the target model's causal structure.

Geiger et al. 2020, 2021

Recipe for causal abstraction

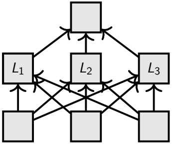
1. State a hypothesis about (an aspect of) the target model's causal structure.
2. Search for an alignment between the causal model and target model.

Geiger et al. 2020, 2021

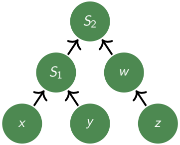
Recipe for causal abstraction

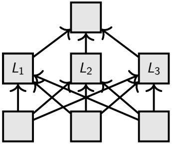
1. State a hypothesis about (an aspect of) the target model's causal structure.
2. Search for an alignment between the causal model and target model.
3. Perform *interchange interventions*.

Geiger et al. 2020, 2021

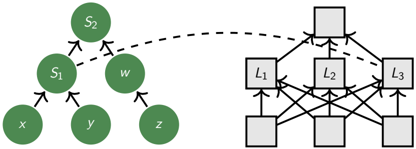


Our neural network successfully adds three numbers.
In human-interpretable terms, how does it do it?

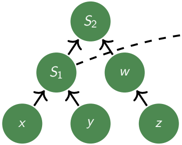


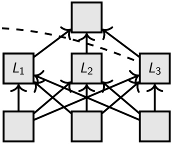


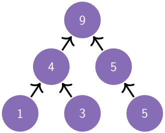
Our causal model adds the first two inputs to form an intermediate variable S_1.



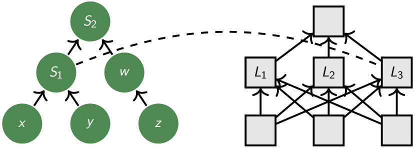
We hypothesize that the neural representation L_3 plays the same role as S_1.

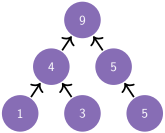


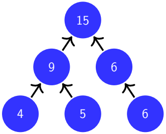




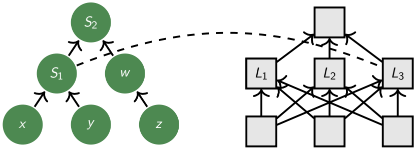
To test this, we run our causal model on $[1, 3, 5]$ and obtain output 9.

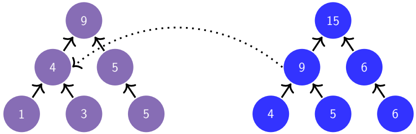




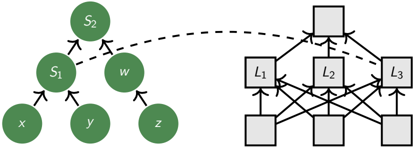


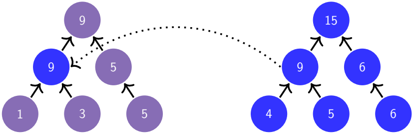
And we run the causal model on [4, 5, 6] to get 15.



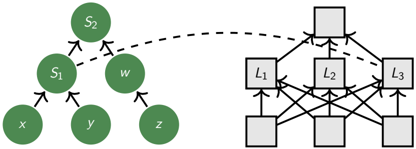


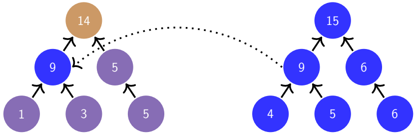
Then we perform an interchange intervention targeting the value of S_1.



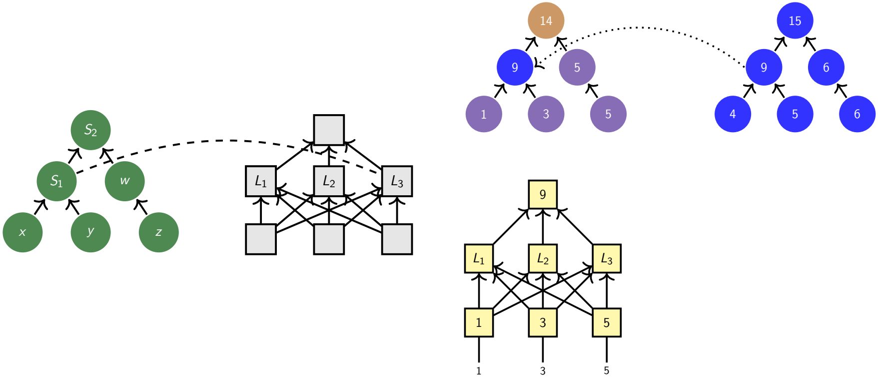


This changes the value of S_1 in the left example to 9.

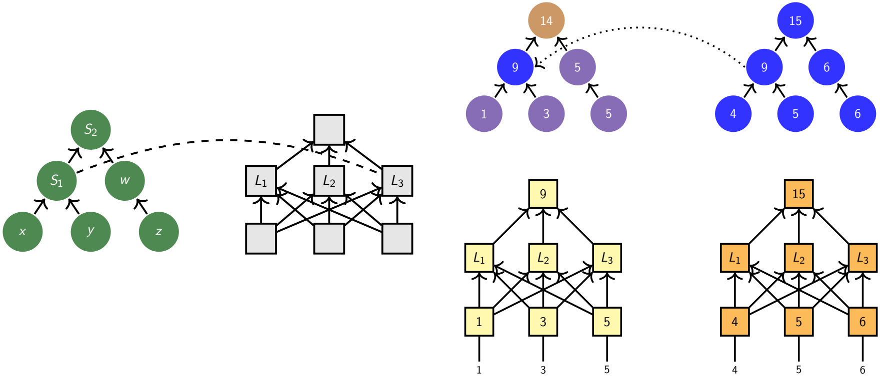




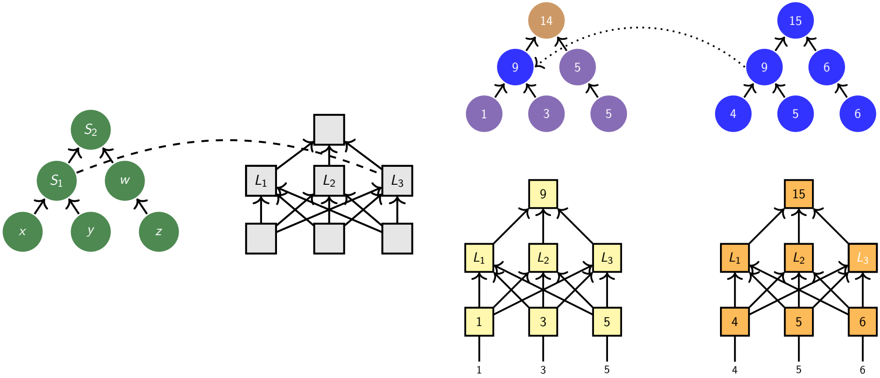
And this causes the model to output 14.



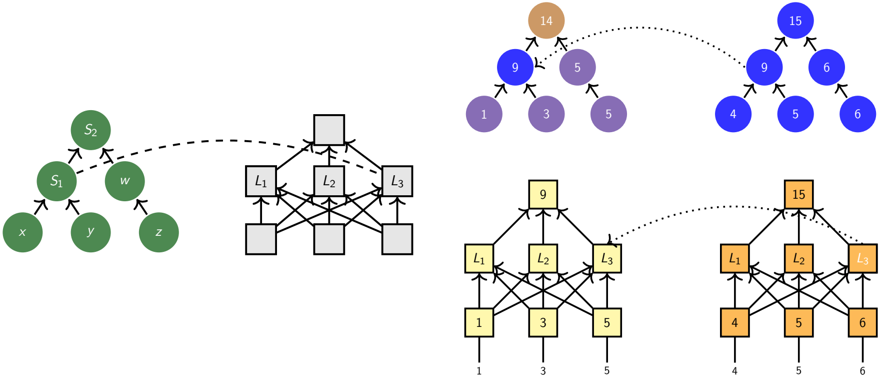
Will the neural network show the same behavior?



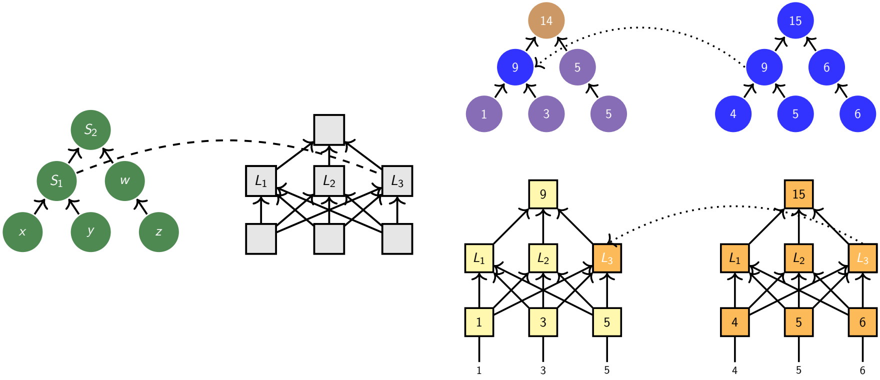
We process the same two examples.



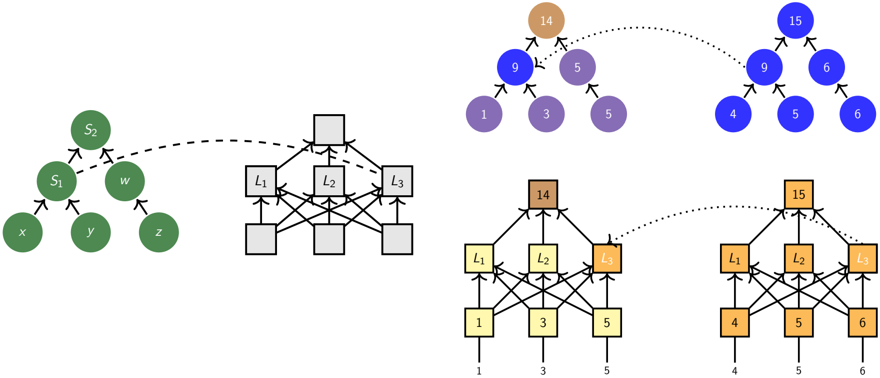
We hypothesized that L_3 plays the role of S_1.



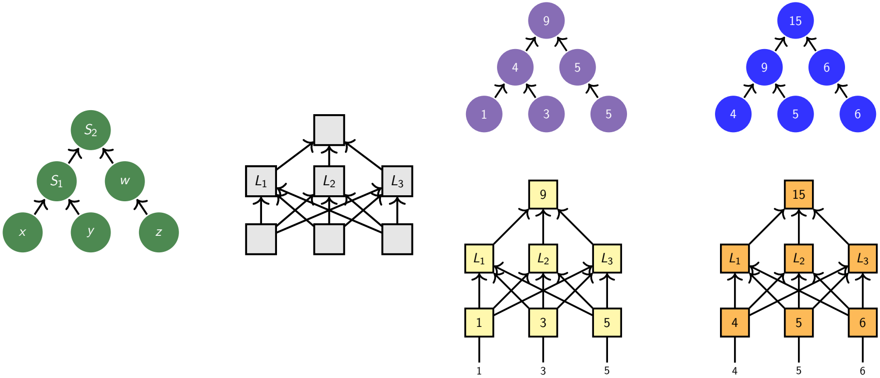
So we perform an intervention targeting L_3.

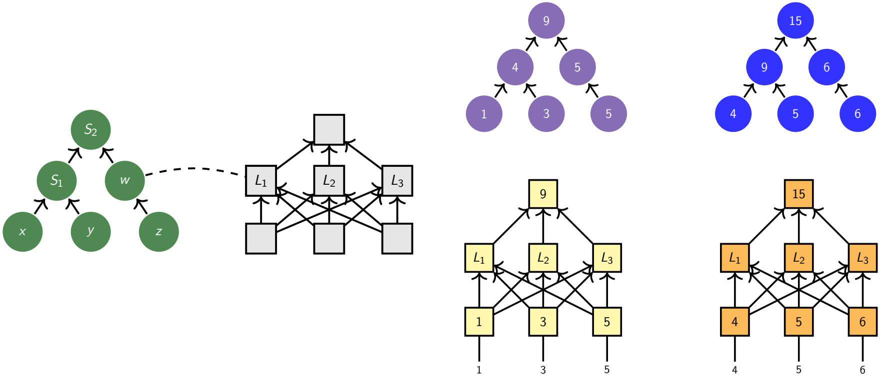


What is the effect of this intervention?

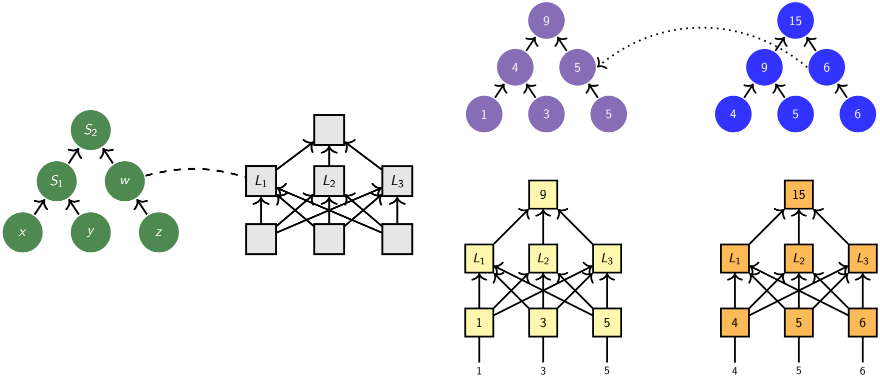


If this leads the network to output 14, we have a piece of evidence that L_3 plays the same role as S_1.

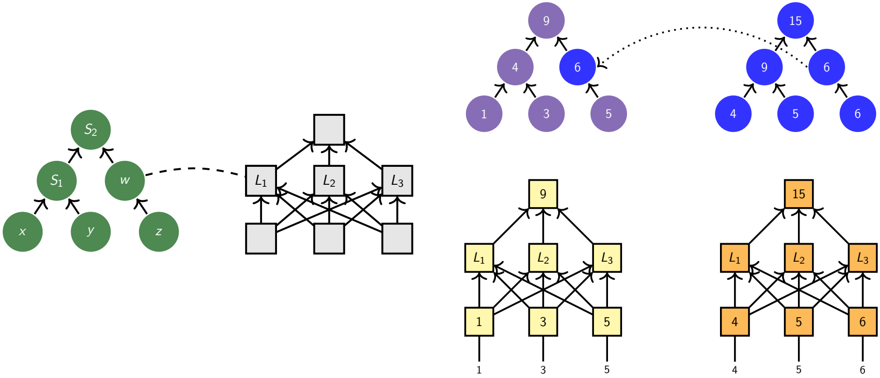

S_2
S_1
w
x
y
z
L_1
L_2
L_3
9
4
5
1
3
5
15
9
6
4
5
6
9
L_1
L_2
L_3
1
3
5
1
3
5
15
L_1
L_2
L_3
4
5
6
4
5
6



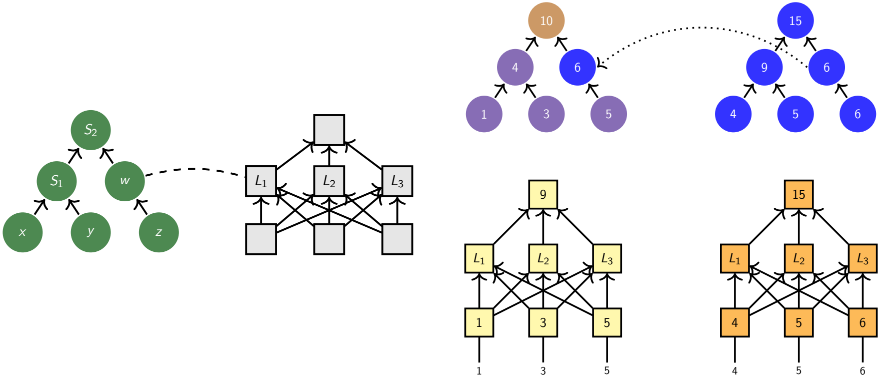
We can repeat the same process using the hypothesis that L_1 plays the role of w.



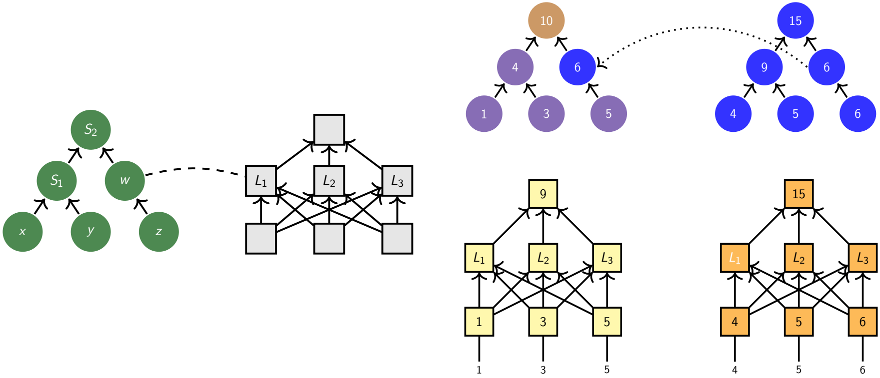
We first intervene on the causal model to get an output for this intervention.



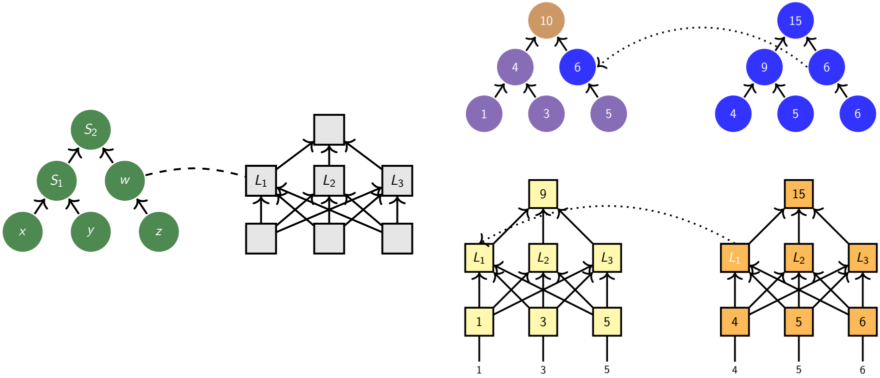
We first intervene on the causal model to get an output for this intervention.



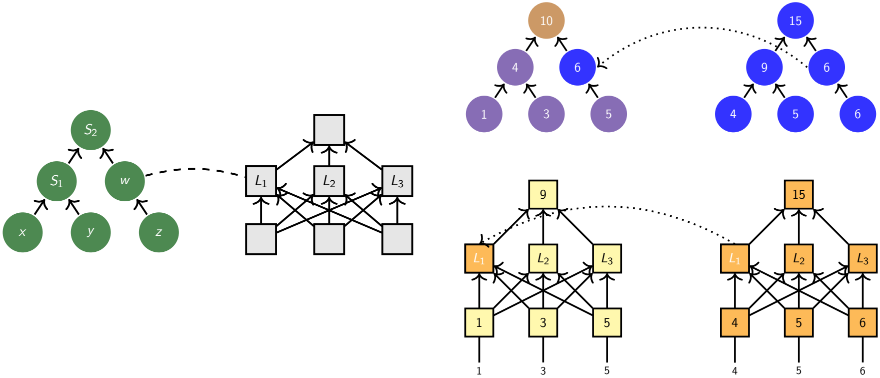
We first intervene on the causal model to get an output for this intervention.



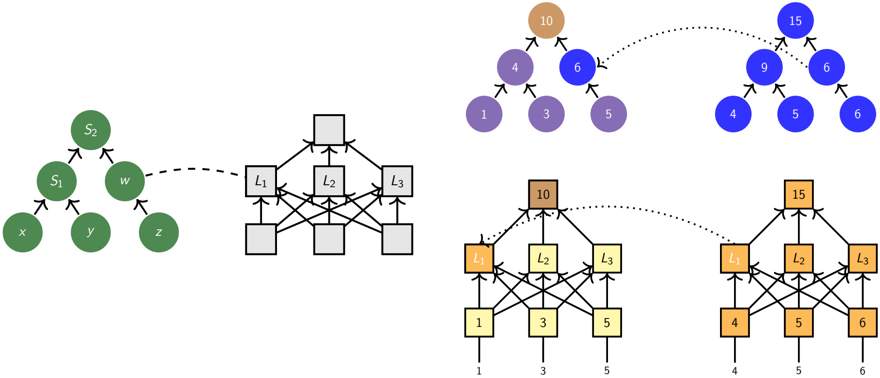
Then we intervene on the neural model.



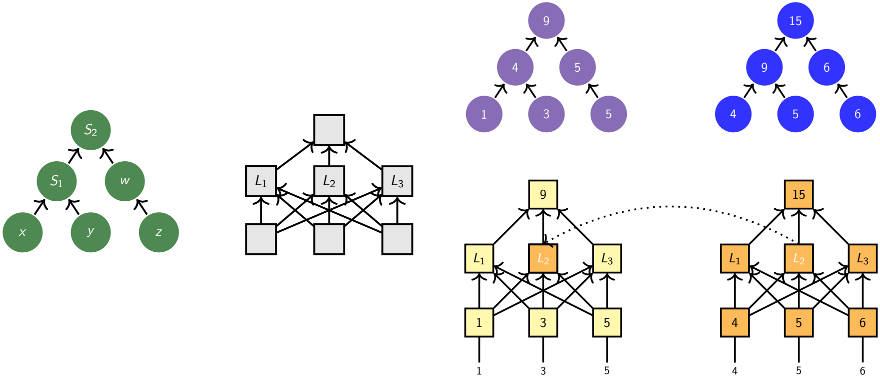
Then we intervene on the neural model.



Then we intervene on the neural model.

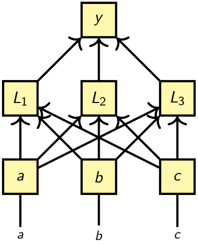


And we check whether the output corresponds to the output of the causal model under the aligned intervention.

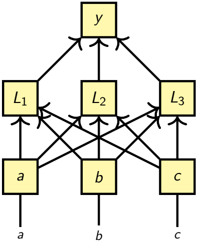


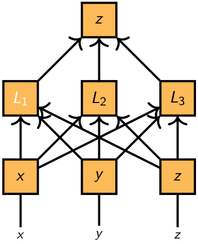
Finally, if we intervene on L_2 and find that the output label never changes, then we have shown that it plays no role in the model's behavior.

Some other interventions

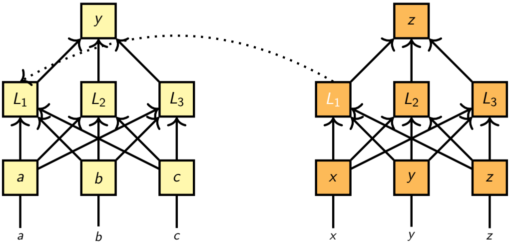


Some other interventions

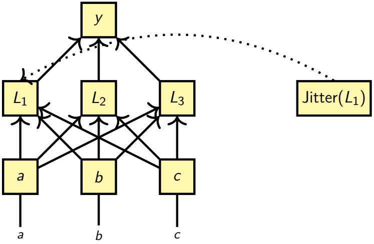




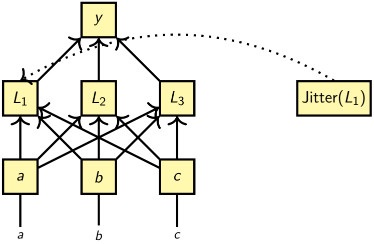
Some other interventions



Some other interventions

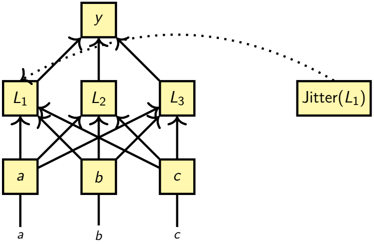


Some other interventions



Potential causal models

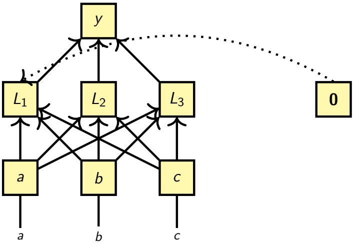
Some other interventions



Potential causal models

- Jitter: Output invariance

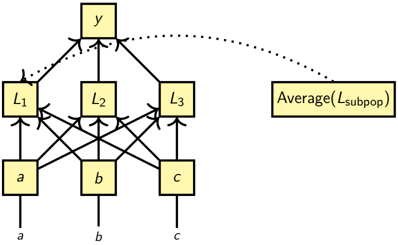
Some other interventions



Potential causal models

- Jitter: Output invariance
- Zero-out: Info removal

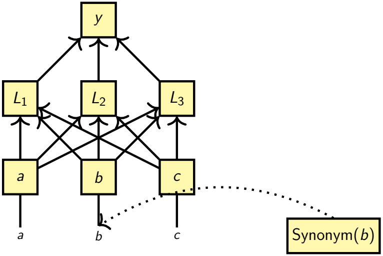
Some other interventions



Potential causal models

- Jitter: Output invariance
- Zero-out: Info removal
- Average vector: Info neutralization

Some other interventions



Potential causal models

- Jitter: Output invariance
- Zero-out: Info removal
- Average vector: Info neutralization
- Data augmentation: Label invariance

Connections to the literature

- Constructive abstraction (Beckers et al. 2020)
- Causal mediation analysis (Vig et al. 2020)
- Role Learning Networks (Soulos et al. 2020)
- CausaLM (Feder et al. 2021)
- Amnesic Probing (Elazar et al. 2021)
- Circuits (Cammarata et al. 2020; Olsson et al. 2022; Wang et al. 2022)
- Causal scrubbing (LawrenceC et al. 2022)

For more:
https://ai.stanford.edu/blog/causal-abstraction/

Findings from causal abstraction

1. Neural networks learn interpretable solutions to hierarchical equality tasks, thereby blurring the distinction between neural and symbolic models (Geiger et al. 2023).
2. Fine-tuned BERT models implement compositional models that allow them to correctly handle hard, out-of-domain natural language inference examples (Geiger et al. 2020, 2021).
3. BART and T5 use coherent entity and situation representations that evolve as the discourse unfolds (Li et al. 2021).

Causal abstraction: Taking stock

1. Verifiably faithful
2. Human interpretable
3. Causal
4. A path to improving models
5. Scalable
6. Minimal assumptions about information encoding

Causal abstraction: Taking stock

1. Verifiably faithful 😃
2. Human interpretable
3. Causal
4. A path to improving models
5. Scalable
6. Minimal assumptions about information encoding

Causal abstraction: Taking stock

1. Verifiably faithful 😃
2. Human interpretable 😃
3. Causal
4. A path to improving models
5. Scalable
6. Minimal assumptions about information encoding

Causal abstraction: Taking stock

1. Verifiably faithful 😃
2. Human interpretable 😃
3. Causal 😃
4. A path to improving models
5. Scalable
6. Minimal assumptions about information encoding

Causal abstraction: Taking stock

1. Verifiably faithful 😃
2. Human interpretable 😃
3. Causal 😃
4. A path to improving models 🤔
5. Scalable
6. Minimal assumptions about information encoding

Causal abstraction: Taking stock

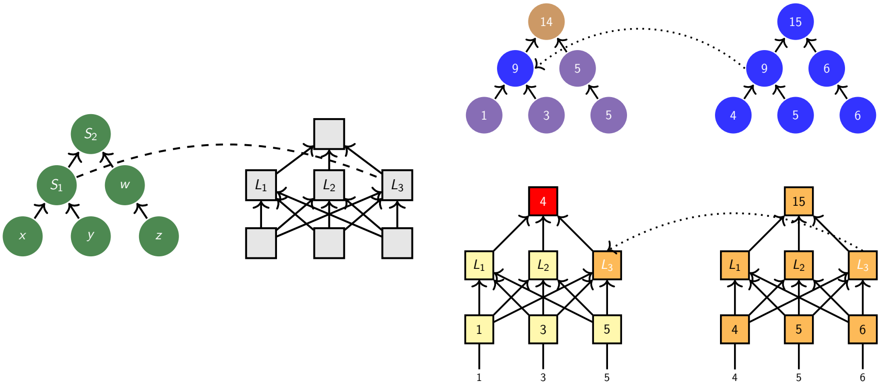
1. Verifiably faithful 😃
2. Human interpretable 😃
3. Causal 😃
4. A path to improving models 🤔
5. Scalable 😟
6. Minimal assumptions about information encoding

Causal abstraction: Taking stock

1. Verifiably faithful 😃
2. Human interpretable 😃
3. Causal 😃
4. A path to improving models 🤔
5. Scalable 😟
6. Minimal assumptions about information encoding 😟

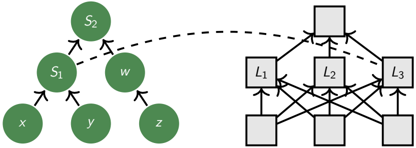
Interchange Intervention Training (IIT)

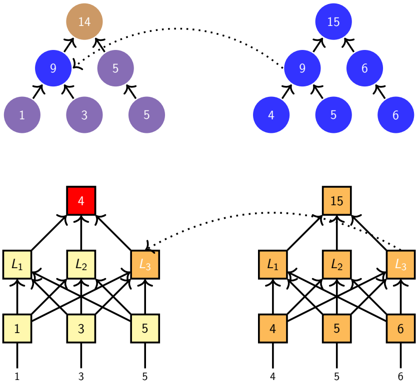
Method



Suppose our network doesn't agree with the causal model under our intervention.

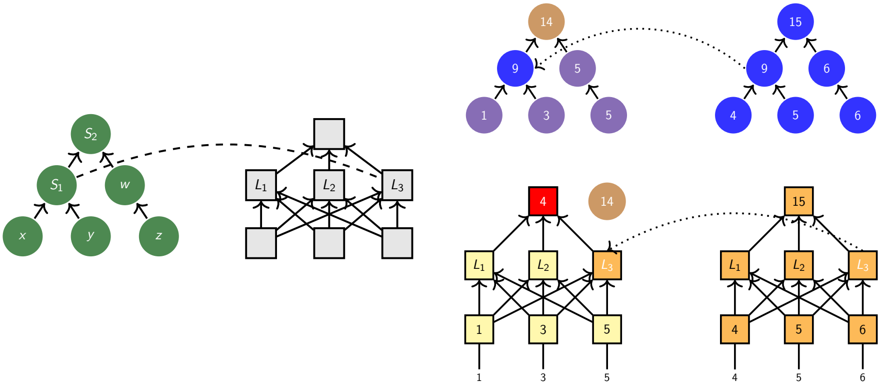
Method





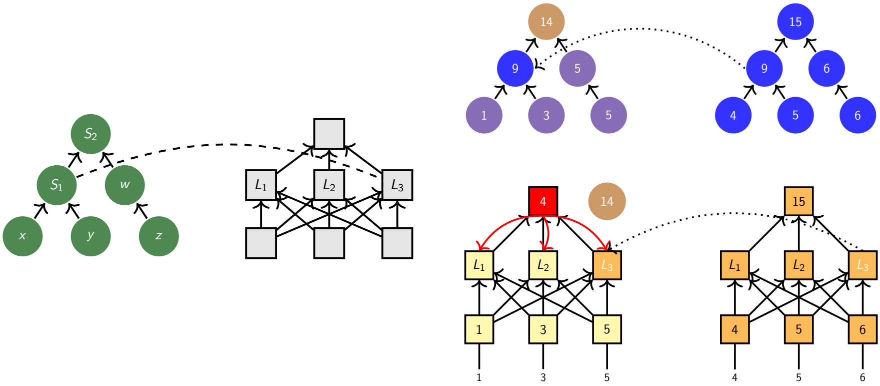
We can correct that misalignment with interchange intervention training.

Method



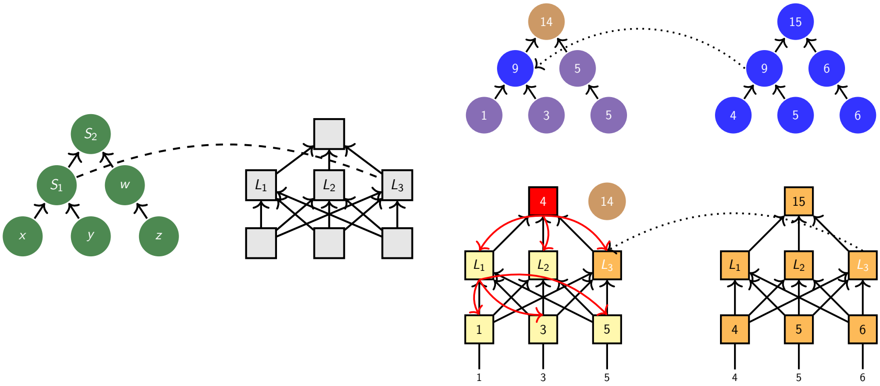
The causal model provides us with a true label, and a comparison with the incorrect prediction gives us an error signal.

Method



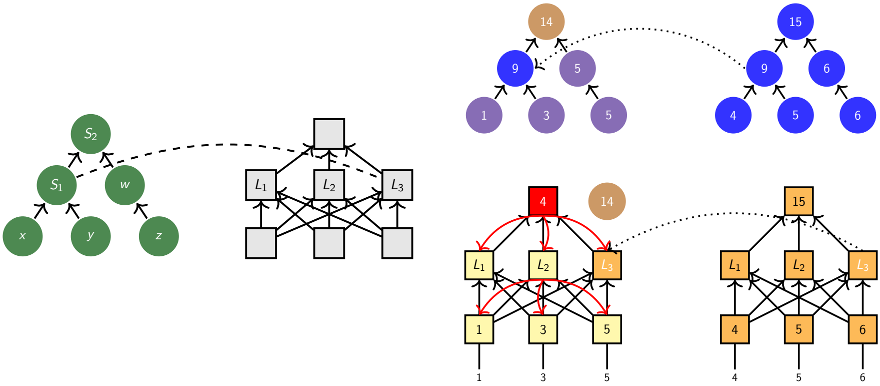
The gradients flow from this node to the top hidden layer as usual.

Method



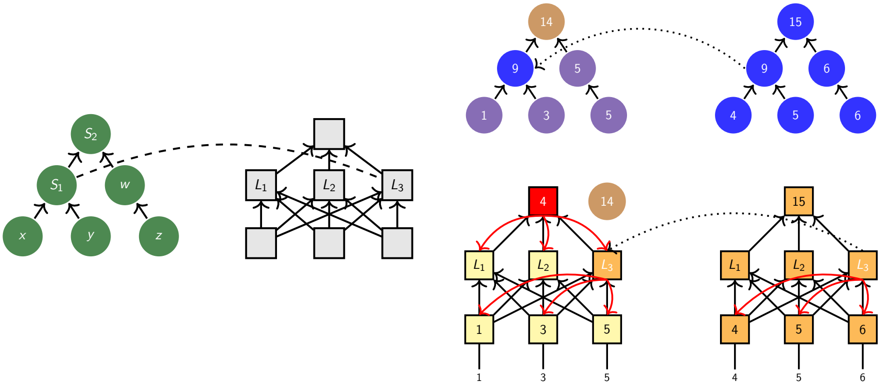
And the gradients flow as usual for the left and center hidden states.

Method



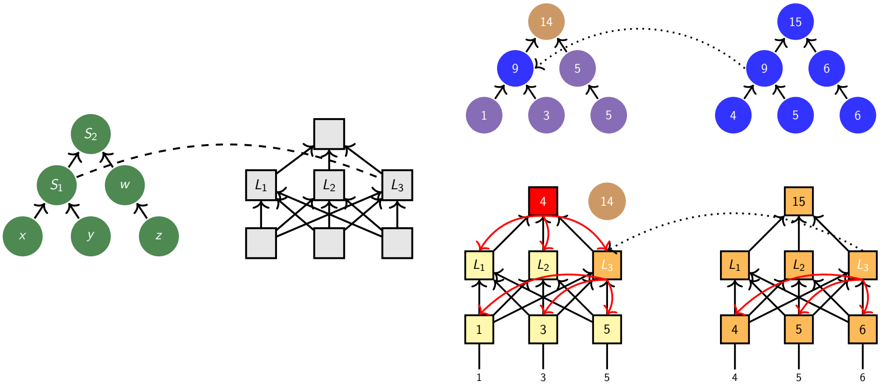
And the gradients flow as usual for the left and center hidden states.

Method



But the intervention site receives a double update, from the target example and the source example at right.

Method



This process gradually brings L_3 into alignment with S_1.

Some applications of IIT

Some applications of IIT

1. Geiger et al. (2022b) develop IIT and use it to achieve state-of-the-art results on the MNIST Pointer Value Retrieval task (MNIST-PVR; Zhang et al. 2021) and the ReaSCAN grounded language understanding benchmark (Wu et al. 2021).

Some applications of IIT

1. Geiger et al. (2022b) develop IIT and use it to achieve state-of-the-art results on the MNIST Pointer Value Retrieval task (MNIST-PVR; Zhang et al. 2021) and the ReaSCAN grounded language understanding benchmark (Wu et al. 2021).
2. Wu et al. (2022) augment the standard distillation objectives (Sanh et al. 2019) with an IIT objective and show that it improves over standard distillation techniques.

Some applications of IIT

1. Geiger et al. (2022b) develop IIT and use it to achieve state-of-the-art results on the MNIST Pointer Value Retrieval task (MNIST-PVR; Zhang et al. 2021) and the ReaSCAN grounded language understanding benchmark (Wu et al. 2021).
2. Wu et al. (2022) augment the standard distillation objectives (Sanh et al. 2019) with an IIT objective and show that it improves over standard distillation techniques.
3. Huang et al. (2023) use IIT to induce internal representations of characters in LMs based in subword tokenization, and they show that this helps with a variety of character-level games and tasks.

Some applications of IIT

1. Geiger et al. (2022b) develop IIT and use it to achieve state-of-the-art results on the MNIST Pointer Value Retrieval task (MNIST-PVR; Zhang et al. 2021) and the ReaSCAN grounded language understanding benchmark (Wu et al. 2021).
2. Wu et al. (2022) augment the standard distillation objectives (Sanh et al. 2019) with an IIT objective and show that it improves over standard distillation techniques.
3. Huang et al. (2023) use IIT to induce internal representations of characters in LMs based in subword tokenization, and they show that this helps with a variety of character-level games and tasks.
4. Wu et al. (2023) use IIT to create concept-level methods for explaining model behavior.

Causal abstraction and IIT: Taking stock

1. Verifiably faithful
2. Human interpretable
3. Causal
4. A path to improving models
5. Scalable
6. Minimal assumptions about information encoding

Causal abstraction and IIT: Taking stock

1. Verifiably faithful 😃
2. Human interpretable 😃
3. Causal 😃
4. A path to improving models
5. Scalable
6. Minimal assumptions about information encoding

Causal abstraction and IIT: Taking stock

1. Verifiably faithful 😃
2. Human interpretable 😃
3. Causal 😃
4. A path to improving models 😃
5. Scalable
6. Minimal assumptions about information encoding

Causal abstraction and IIT: Taking stock

1. Verifiably faithful 😃
2. Human interpretable 😃
3. Causal 😃
4. A path to improving models 😃
5. Scalable 😟
6. Minimal assumptions about information encoding 😟

Boundless Distributed Alignment Search (DAS)

Our scorecard again

1. Verifiably faithful 😃
2. Human interpretable 😃
3. Causal 😃
4. A path to improving models 😃
5. Scalable:
6. Minimal assumptions about information encoding:

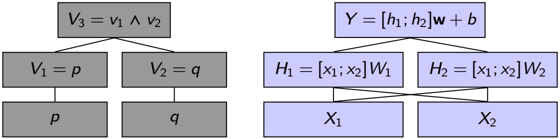
Our scorecard again

1. Verifiably faithful 😃
2. Human interpretable 😃
3. Causal 😃
4. A path to improving models 😃
5. Scalable: Alignment search is expensive.
6. Minimal assumptions about information encoding:

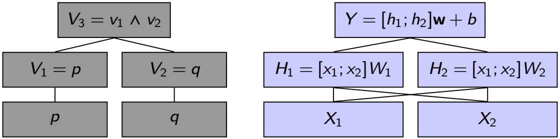
Our scorecard again

1. Verifiably faithful 😃
2. Human interpretable 😃
3. Causal 😃
4. A path to improving models 😃
5. Scalable: Alignment search is expensive.
6. Minimal assumptions about information encoding: We search only in a standard basis and assume groups of neurons will play unique roles.

A simple causal abstraction analysis



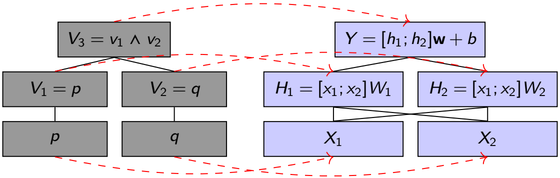
A simple causal abstraction analysis



$$W_1 = \begin{bmatrix} \cos(20°) & -\sin(20°) \end{bmatrix} \qquad \mathbf{w} = \begin{bmatrix} 1 & 1 \end{bmatrix}$$

$$W_2 = \begin{bmatrix} \sin(20°) & \cos(20°) \end{bmatrix} \qquad b = -1.8$$

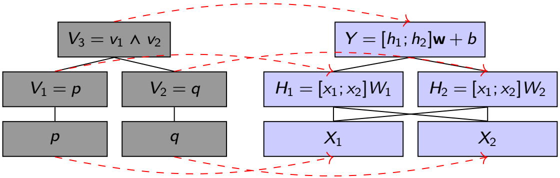
A simple causal abstraction analysis



$$W_1 = \begin{bmatrix} \cos(20°) & -\sin(20°) \end{bmatrix} \qquad \mathbf{w} = \begin{bmatrix} 1 & 1 \end{bmatrix}$$

$$W_2 = \begin{bmatrix} \sin(20°) & \cos(20°) \end{bmatrix} \qquad b = -1.8$$

A simple causal abstraction analysis



$$W_1 = \begin{bmatrix} \cos(20°) & -\sin(20°) \end{bmatrix} \qquad \mathbf{w} = \begin{bmatrix} 1 & 1 \end{bmatrix}$$

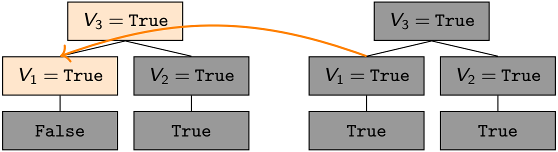
$$W_2 = \begin{bmatrix} \sin(20°) & \cos(20°) \end{bmatrix} \qquad b = -1.8$$

The high-level model **does not abstract** the new neural model under our chosen alignment.

Interchange intervention failure

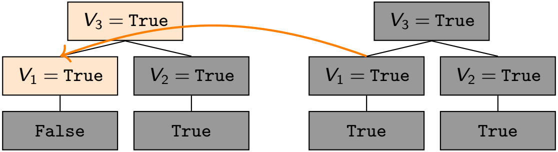
Interchange intervention failure

An interchange intervention on the high-level model:

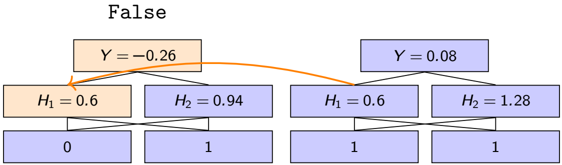


Interchange intervention failure

An interchange intervention on the high-level model:

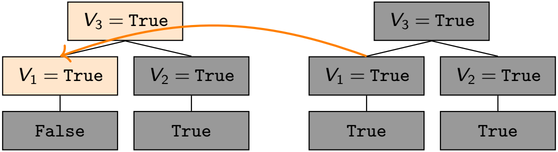


The aligned interchange intervention on the neural model:

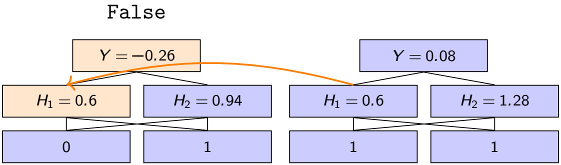


Interchange intervention failure

An interchange intervention on the high-level model:

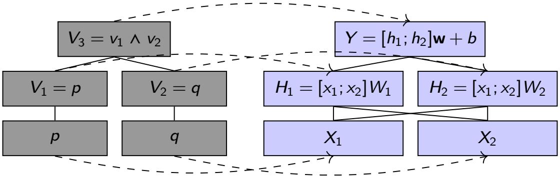


The aligned interchange intervention on the neural model:



The two models have **unequal counterfactual predictions**

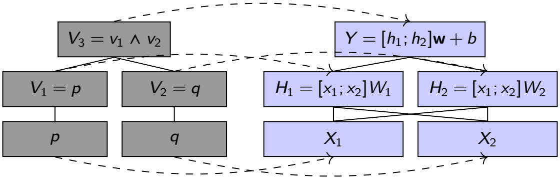
But the relationship holds in a non-standard basis



$$W_1 = \begin{bmatrix} \cos(20°) & -\sin(20°) \end{bmatrix} \qquad \mathbf{w} = \begin{bmatrix} 1 & 1 \end{bmatrix}$$

$$W_2 = \begin{bmatrix} \sin(20°) & \cos(20°) \end{bmatrix} \qquad b = -1.8$$

But the relationship holds in a non-standard basis



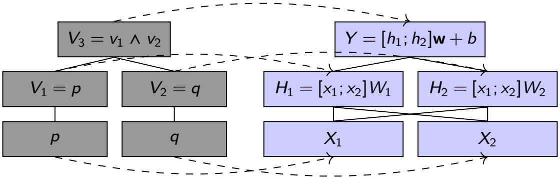
$$W_1 = \begin{bmatrix} \cos(20°) & -\sin(20°) \end{bmatrix} \qquad \mathbf{w} = \begin{bmatrix} 1 & 1 \end{bmatrix}$$

$$W_2 = \begin{bmatrix} \sin(20°) & \cos(20°) \end{bmatrix} \qquad b = -1.8$$

View $[H_1, H_2]$ under a non-standard basis by rotating $-20°$:

$$\begin{bmatrix} \cos(-20°) & -\sin(-20°) \\ \sin(-20°) & \cos(-20°) \end{bmatrix}$$

But the relationship holds in a non-standard basis



$$W_1 = \begin{bmatrix} \cos(20°) & -\sin(20°) \end{bmatrix} \qquad \mathbf{w} = \begin{bmatrix} 1 & 1 \end{bmatrix}$$

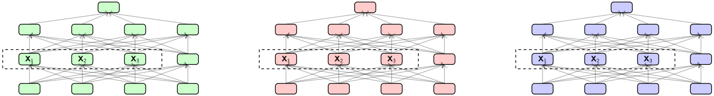
$$W_2 = \begin{bmatrix} \sin(20°) & \cos(20°) \end{bmatrix} \qquad b = -1.8$$

View $[H_1, H_2]$ under a non-standard basis by rotating $-20°$:

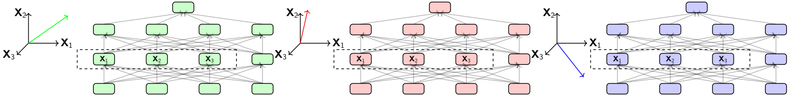
$$\begin{bmatrix} \cos(-20°) & -\sin(-20°) \\ \sin(-20°) & \cos(-20°) \end{bmatrix}$$

Boundless DAS: Freeze the target model parameters and learn a rotation matrix and the boundaries of the intervention to maximize interchange intervention accuracy.

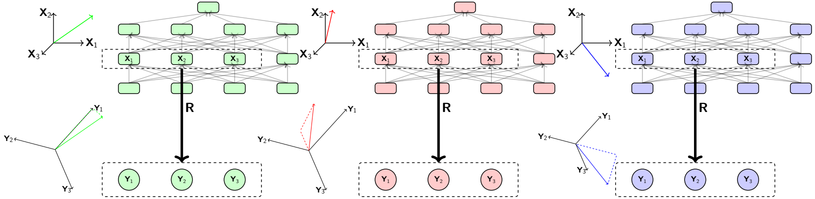
Solution: Distributed Interchange Intervention



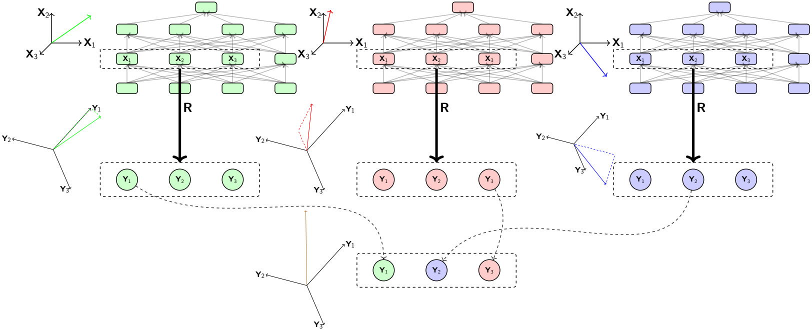
Solution: Distributed Interchange Intervention



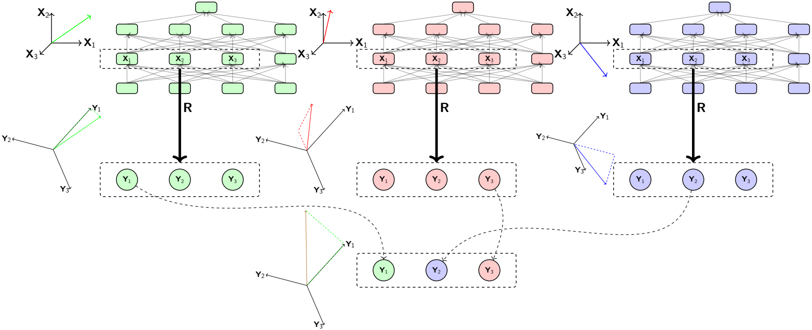
Solution: Distributed Interchange Intervention



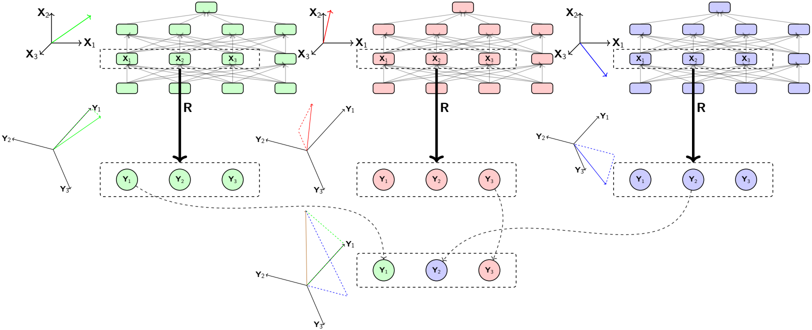
Solution: Distributed Interchange Intervention



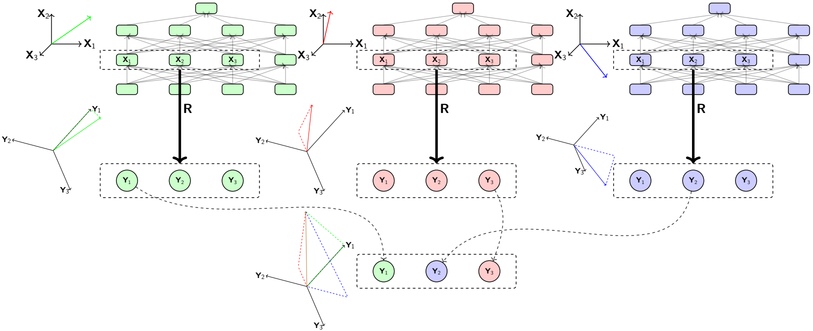
Solution: Distributed Interchange Intervention



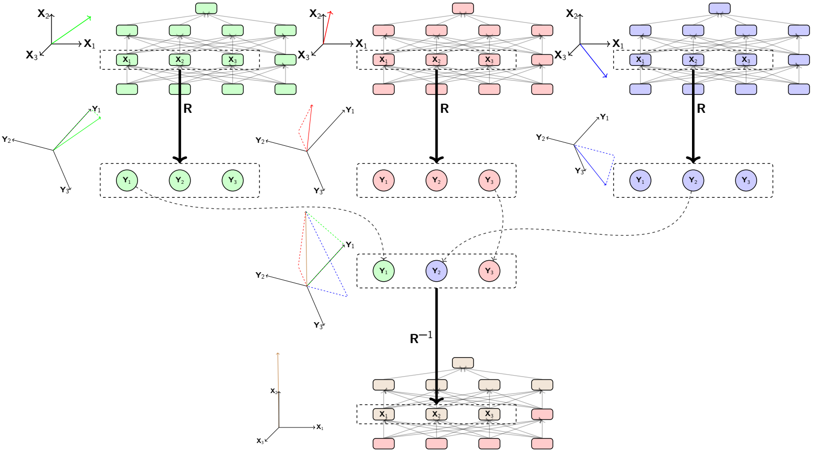
Solution: Distributed Interchange Intervention



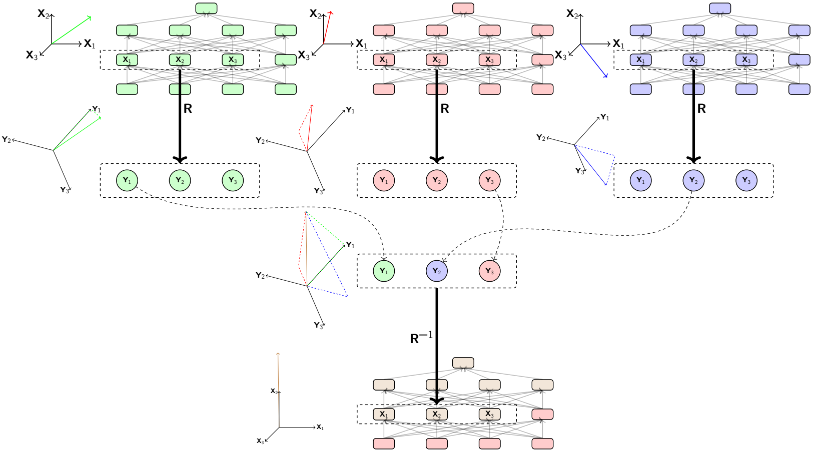
Solution: Distributed Interchange Intervention



Solution: Distributed Interchange Intervention



Solution: Distributed Interchange Intervention



Freeze the model parameters and **learn** a rotation matrix with distributed interchange intervention training as well as the boundaries of the intervention.

Boundless DAS: Taking stock

1. Verifiably faithful
2. Human interpretable
3. Causal
4. A path to improving models
5. Scalable
6. Minimal assumptions about information encoding

Boundless DAS: Taking stock

1. Verifiably faithful 😃
2. Human interpretable 😃
3. Causal 😃
4. A path to improving models 😃
5. Scalable
6. Minimal assumptions about information encoding

Boundless DAS: Taking stock

1. Verifiably faithful 😃
2. Human interpretable 😃
3. Causal 😃
4. A path to improving models 😃
5. Scalable 😃
6. Minimal assumptions about information encoding

Boundless DAS: Taking stock

1. Verifiably faithful 😃
2. Human interpretable 😃
3. Causal 😃
4. A path to improving models 😃
5. Scalable 😃
6. Minimal assumptions about information encoding 😃

Identifying Causal Mechanisms in Alpaca

Price Tagging Game

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Please say yes only if it costs between **[X.XX]** and **[X.XX]** dollars, otherwise no.

Input:
[X.XX]

Response:
[Model Output]

Identifying Causal Mechanisms in Alpaca

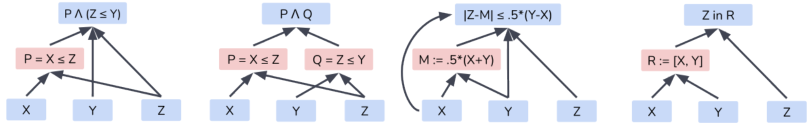
Price Tagging Game

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Please say yes only if it costs between **[X.XX]** and **[X.XX]** dollars, otherwise no.

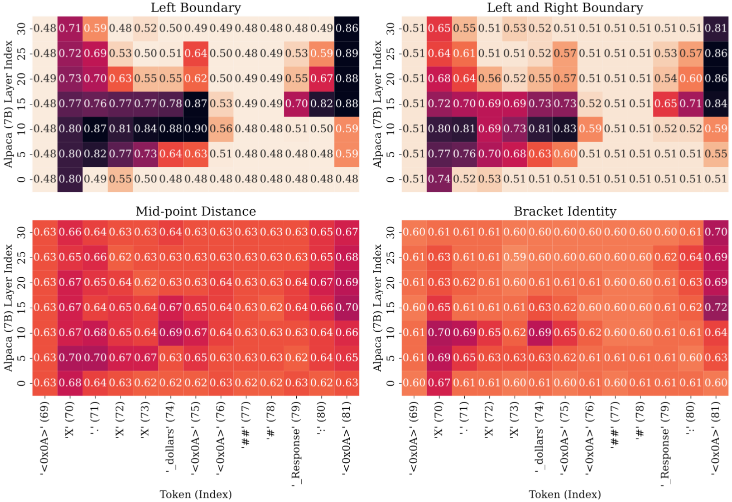
Input:
[X.XX]

Response:
[Model Output]

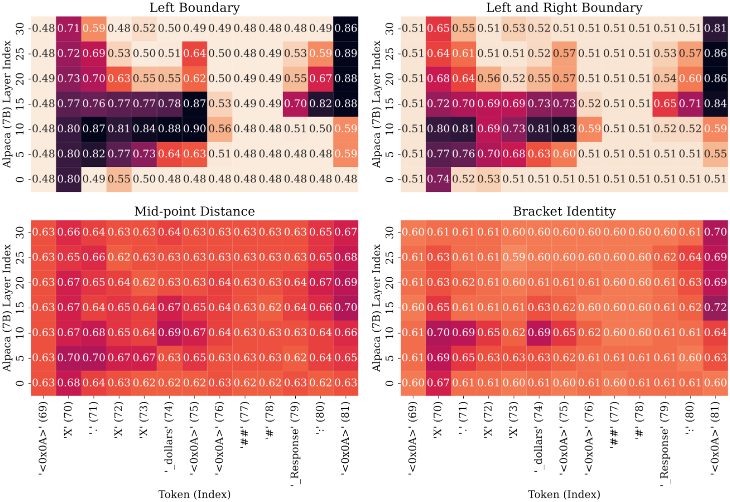


Left Boundary | **Left and Right Boundary** | **Mid-point Distance** | **Bracket Identity**

Identifying Causal Mechanisms in Alpaca

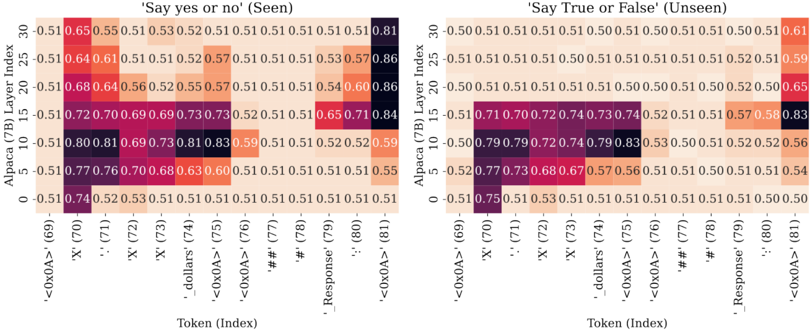


Identifying Causal Mechanisms in Alpaca



Learned DAS solution transfers to many variations of the input instructions, and even the output space.

Identifying Causal Mechanisms in Alpaca



Conclusions

Reminder: A crucial prerequisite

Reminder: A crucial prerequisite

Identify approved uses

Reminder: A crucial prerequisite

Safety

Identify
approved uses

Reminder: A crucial prerequisite

Trust

Safety

Identify approved uses

Reminder: A crucial prerequisite

Trust

Safety

Mitigate social biases

Identify approved uses

Reminder: A crucial prerequisite

Trust

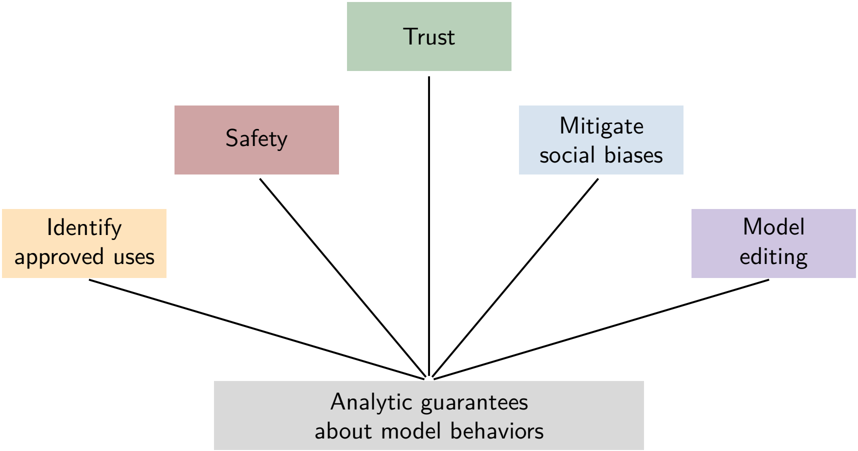
Safety

Mitigate social biases

Identify approved uses

Model editing

Reminder: A crucial prerequisite



The near future of explainability research

The near future of explainability research

1. Deeper causal explanations

The near future of explainability research

1. Deeper causal explanations
2. Human-interpretable explanations

The near future of explainability research

1. Deeper causal explanations
2. Human-interpretable explanations
3. Automatic discovery of causal models

The near future of explainability research

1. Deeper causal explanations
2. Human-interpretable explanations
3. Automatic discovery of causal models
4. Applications to ever-larger foundation models

The near future of explainability research

1. Deeper causal explanations
2. Human-interpretable explanations
3. Automatic discovery of causal models
4. Applications to ever-larger foundation models
5. Increasing evidence that models are inducing relevant causal structure about our world.

The near future of explainability research

1. Deeper causal explanations
2. Human-interpretable explanations
3. Automatic discovery of causal models
4. Applications to ever-larger foundation models
5. Increasing evidence that models are inducing relevant causal structure about our world.

Thanks!

References I

Sander Beckers, Frederick Eberhardt, and Joseph Y. Halpern. 2020. Approximate causal abstractions. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 606–615. PMLR.

Nick Cammarata, Shan Carter, Gabriel Goh, Chris Olah, Michael Petrov, Ludwig Schubert, Chelsea Voss, Ben Egan, and Swee Kiat Lim. 2020. Thread: Circuits. *Distill*. Https://distill.pub/2020/circuits.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9(0):160–175.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2021. CausaLM: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 9574–9586.

Atticus Geiger, Christopher Potts, and Thomas Icard. 2023. Causal abstraction for faithful model interpretation. Ms., Stanford University.

Atticus Geiger, Kyle Richardson, and Christopher Potts. 2020. Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 163–173, Online. Association for Computational Linguistics.

Atticus Geiger, Zhengxuan Wu, Karel D'Oosterlinck, Elisa Kreiss, Noah D. Goodman, Thomas Icard, and Christopher Potts. 2022a. Faithful, interpretable model explanations via causal abstraction. Stanford AI Lab Blog.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. 2022b. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 7324–7338. PMLR.

Jing Huang, Zhengxuan Wu, Kyle Mahowald, and Christopher Potts. 2023. Inducing character-level structure in subword-based language models with type-level interchange intervention training. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12163–12180, Toronto, Canada. Association for Computational Linguistics.

LawrenceC, Adrià Garriga-alonso, Nicholas Goldowsky-Dill, ryan_greenblatt, jenny, Ansh Radhakrishnan, Buck, and Nate Thomas. 2022. Causal scrubbing: a method for rigorously testing interpretability hypotheses. Blog post, Redwood Research.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827, Online. Association for Computational Linguistics.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108.

References II

Paul Soulos, R. Thomas McCoy, Tal Linzen, and Paul Smolensky. 2020. Discovering the compositional structure of vector representations with role learning networks. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 238–254, Online. Association for Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural nlp: The case of gender bias.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*.

Zhengxuan Wu, Karel D'Oosterlinck, Atticus Geiger, Amir Zur, and Christopher Potts. 2023. Causal Proxy Models for concept-based model explanations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 37313–37334. PMLR.

Zhengxuan Wu, Atticus Geiger, Josh Rozner, Elisa Kreiss, Hanson Lu, Thomas Icard, Christopher Potts, and Noah D. Goodman. 2022. Causal distillation for language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4288–4295, Seattle, United States. Association for Computational Linguistics.

Zhengxuan Wu, Elisa Kreiss, Desmond C. Ong, and Christopher Potts. 2021. ReaSCAN: Compositional reasoning in language grounding. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Chiyuan Zhang, Maithra Raghu, Jon M. Kleinberg, and Samy Bengio. 2021. Pointer value retrieval: A new benchmark for understanding the limits of neural network generalization. *CoRR*, abs/2107.12580.