# Causal Foundations for Interpretability

## What Your Methods Actually Identify


| | |
|---|---|
| Shruti Joshi | Mila, Université de Montréal |
| Aaron Mueller | Boston University |
| Patrik Reizinger | Mila, Université de Montréal & ELLIS |
| Dhanya Sridhar | Mila, Université de Montréal |



EMNLP 2026 · Budapest
Wednesday 28 October 2026, 14:00–17:30 · causality4mechinterp.github.io

# About Us



Shruti Joshi

PhD Student

Mila, Université de Montréal



Aaron Mueller

Assistant Professor

Boston University



Patrik Reizinger

Postdoctoral Fellow

Mila, Université de Montréal, ELLIS



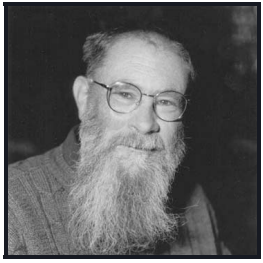
Dhanya Sridhar

Assistant Professor

Mila, Université de Montréal

# What is causation?

# What is causation?

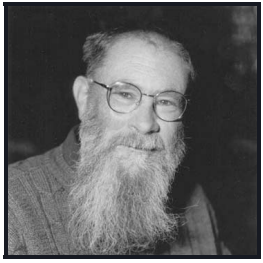


David Lewis

Where *C* and *E* are two distinct possible events, *E* causally depends on *C* if and only if, if *C* were to occur *E* would occur, and if *C* were not to occur *E* would not occur.

—"Causation", *Journal of Philosophy*, 1973

# What is causation?



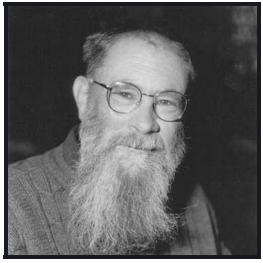
David Lewis

Where *C* and *E* are two distinct possible events, *E* causally depends on *C* if and only if, if *C* were to occur *E* would occur, and if *C* were not to occur *E* would not occur.

—"Causation", *Journal of Philosophy*, 1973

This theory also supports **causal chains** from *C* to *E* (rather than direct links).

# What is causation?



David Lewis

Where *C* and *E* are two distinct possible events, *E* causally depends on *C* if and only if, if *C* were to occur *E* would occur, and if *C* were not to occur *E* would not occur.

—"Causation", *Journal of Philosophy*, 1973

This theory also supports **causal chains** from *C* to *E* (rather than direct links).

Moreover, we can speak of counterfactual *influence* on the way in which the event occurs, rather than just whether or not it happens at all.

# What is interpretability?

Interpretability asks **how** and **why** questions:

**How** (by what means) did my model do that?

**Why** did my model do that?

This stands in contrast to **what** and **can** questions: **What** is my model doing (e.g., benchmarking, behavioral analysis)? **Can** models do this?

# Why should interpretability care about causality?

- Causal evidence supports far stronger claims than correlational evidence.
- We are lucky in that causal evidence is shockingly easy to obtain relative to other fields. We should take advantage of this!
- Many interpretability claims are *implicitly* causal claims, even if not presented as such.

# Interpretability questions

# Interpretability questions

- Does an LLM use concept X when it outputs tokens correlated with X?

# Interpretability questions

- Does an LLM use concept X when it outputs tokens correlated with X?
- Where does an LLM represent a fact (e.g., the Eiffel Tower being in Paris)?

# Interpretability questions

- Does an LLM use concept X when it outputs tokens correlated with X?
- Where does an LLM represent a fact (e.g., the Eiffel Tower being in Paris)?
- Do LLM activations represent a concept via a linear direction in activation space?

# Interpretability questions

- Does an LLM use concept X when it outputs tokens correlated with X?
- Where does an LLM represent a fact (e.g., the Eiffel Tower being in Paris)?
- Do LLM activations represent a concept via a linear direction in activation space?
- Does an LLM add two numbers by adding the ones place and tens place separately, then carrying the one if needed?

# Interpretability questions

- Does an LLM use concept X when it outputs tokens correlated with X?
- Where does an LLM represent a fact (e.g., the Eiffel Tower being in Paris)?
- Do LLM activations represent a concept via a linear direction in activation space?
- Does an LLM add two numbers by adding the ones place and tens place separately, then carrying the one if needed?

*These are causal questions!*

Does the LLM use concept X when it outputs tokens correlated with X?



If (the representation of) X changed and nothing else, would the output have changed?

(I.e., is the output counterfactually dependent on (the representation of) X?)

# Outline

## Asking causal questions

causal mediation

## Levels of causal questions

the causal ladder

## Identifying the answers

when are answers to causal questions unique?

# Notation

| | |
|---|---|
| $x$ | input |
| $x'$ | an input distinct from $x$ |
| $\ell$ | a model layer |
| $a$ | an activation at a given token position |
| $y$ | a model or probe output |
| $m$ | some metric (usually derived from $y$) |
| $\text{do}(a := \tilde{a})$ | an intervention where $a$ is manually set to $\tilde{a}$ |
| $\tilde{y}$ | the model/probe output after rerunning with an intervention |
| $\varphi$ | a featurizer |

# Outline

## Asking causal questions

causal mediation

## Levels of causal questions

the causal ladder

## Identifying the answers

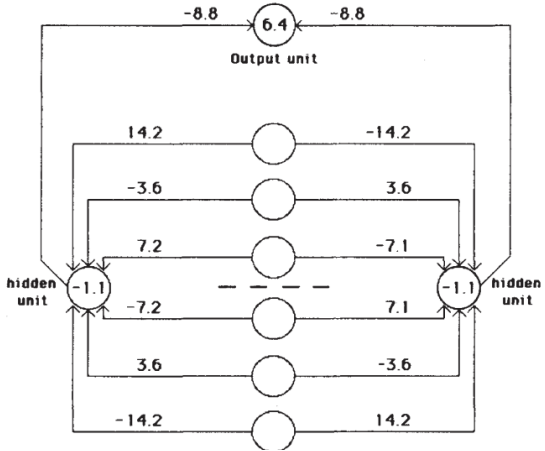
when are answers to causal questions unique?

# Outline

**Asking causal questions** Aaron Mueller
causal mediation

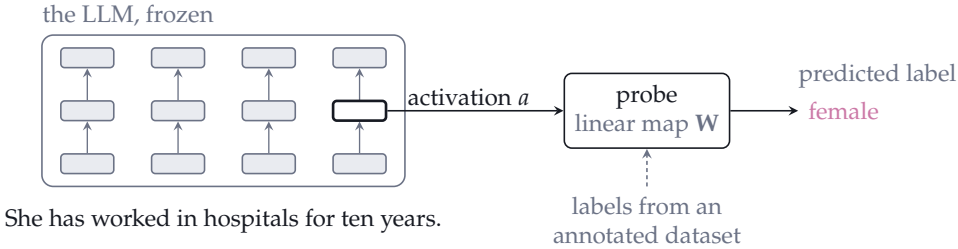
Levels of causal questions
the causal ladder

Identifying the answers
when are answers to causal questions unique?




Rumelhart et al. (1986)

# Probing


the LLM, frozen
activation $a$
probe
linear map $\mathbf{W}$
predicted label
female
She has worked in hospitals for ten years.
labels from an
annotated dataset

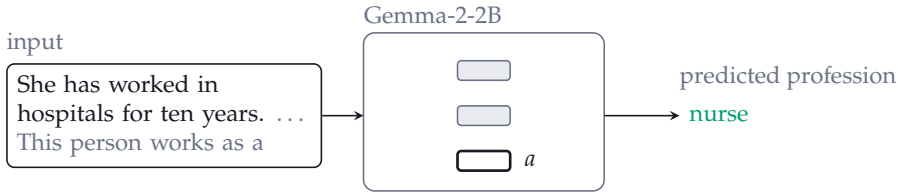
# Probing

**Probes** are classifiers trained on the activations of neural networks. They aim to locate representations of particular concepts:

- Parts of speech — Hewitt and Liang (2019)
- Dependency tree distances — Hewitt and Manning (2019)
- The truthfulness of a statement — Marks and Tegmark (2024)
- Reward hacking behaviors — Bergen et al. (2026)

# Probing

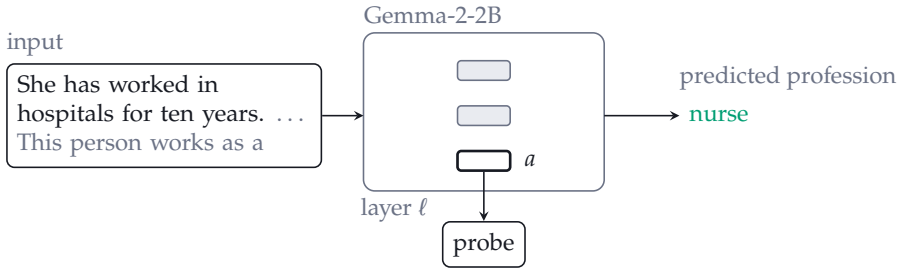
Let's start by trying to probe for gender representations:




Biographies from Bias in Bios (De-Arteaga et al., 2019)

# Probing

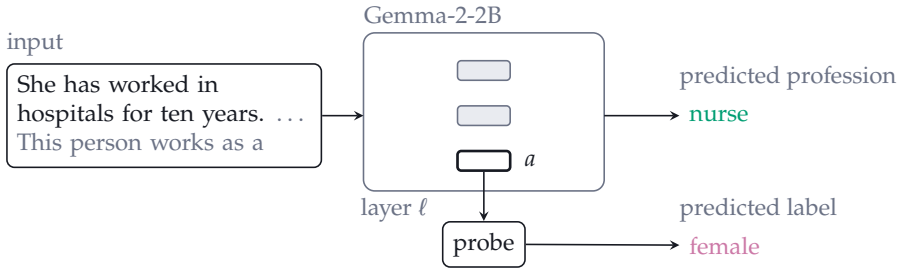
Let's start by trying to probe for gender representations:




Suppose a linear probe on $a$ reads gender with 100% accuracy.

# Probing

Let's start by trying to probe for gender representations:




Suppose a linear probe on $a$ reads gender with 100% accuracy. It would be easy to conclude at first glance that the model uses gender information.

Biographies from Bias in Bios (De-Arteaga et al., 2019)

# Interpreting probing results

## Target question

How much does the prediction
change when gender changes, via $a$?

# Interpreting probing results

Target question

How much does the prediction change when gender changes, via $a$?

Actual measurement

How well a linear map of $a$ predicts gender on this particular data.

# Interpreting probing results

| Target question | Actual measurement |
| --- | --- |
| How much does the prediction change when gender changes, via $a$? | How well a linear map of $a$ predicts gender on this particular data. |

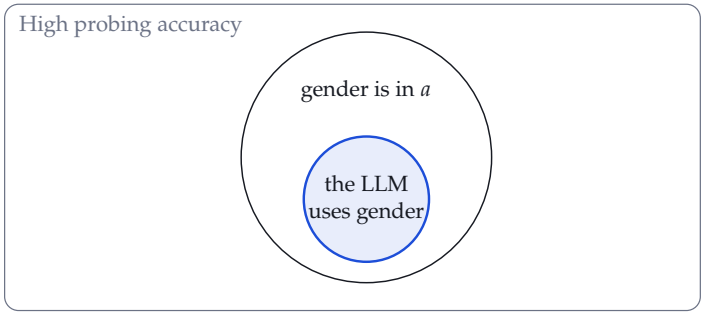
**Estimand** A quantity that would answer the question of interest, if it can be computed exactly.

# Interpreting probing results

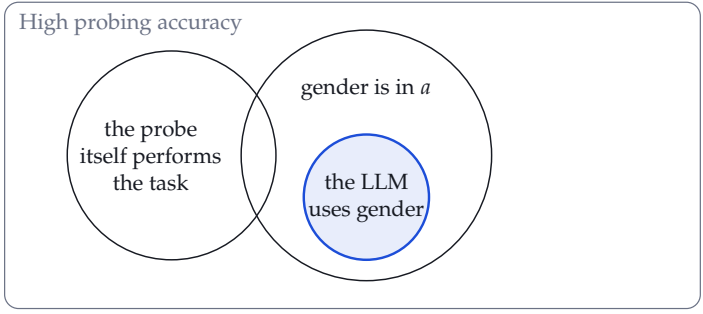
| Target question | Actual measurement |
| --- | --- |
| How much does the prediction change when gender changes, via $a$? | How well a linear map of $a$ predicts gender on this particular data. |

| | |
| --- | --- |
| Estimand | A quantity that would answer the question of interest, if it can be computed exactly. |
| Estimator | A procedure approximating the estimand from data. |

# Interpreting probing results

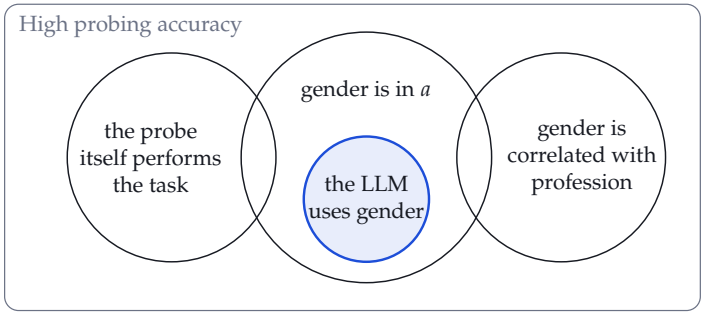


# Interpreting probing results



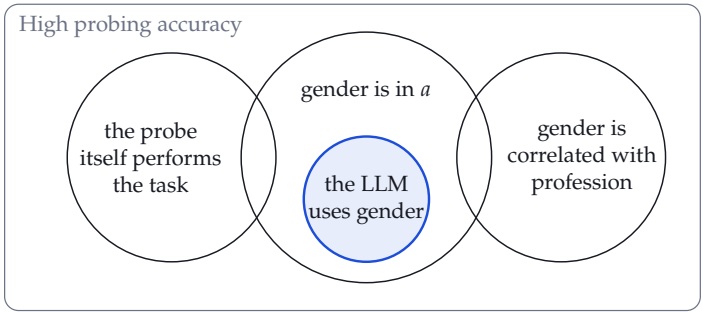

As tested by: a control task

# Interpreting probing results



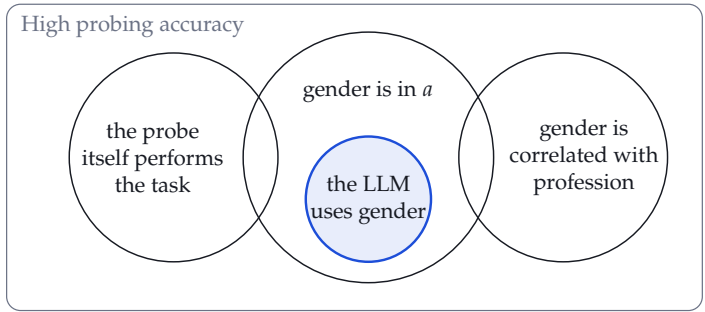

As tested by: a control task randomizing gender

# Interpreting probing results



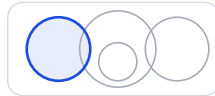

As tested by: a control task | changing *a* | randomizing gender

# Interpreting probing results

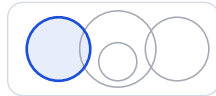



Only the inner circle answers our target question.

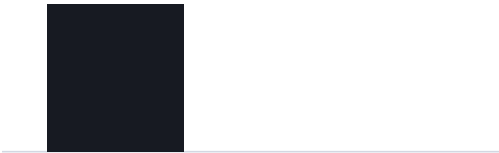
# Control tasks



# Control tasks



ELMo layer 1, linear probe

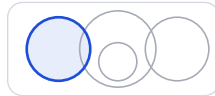


Part of speech, 97.2%

Hewitt and Liang (2019), Table 1

Demo: Shuffled probe

# Control tasks



ELMo layer 1, linear probe

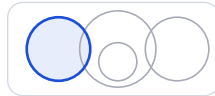
Selectivity, 26.0 points

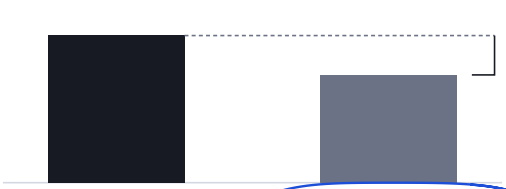



Hewitt and Liang (2019), Table 1

Demo: Shuffled probe

# Control tasks

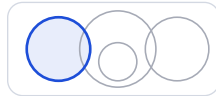


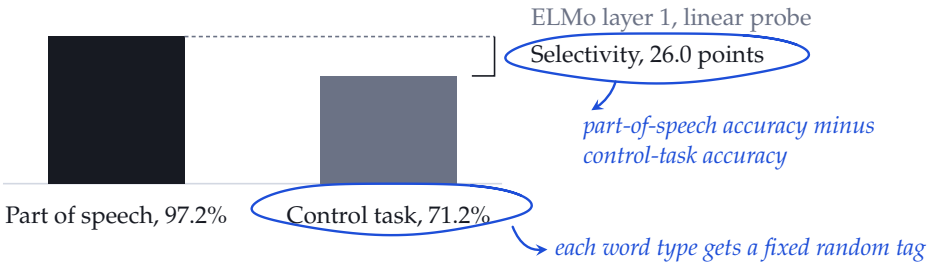



Hewitt and Liang (2019), Table 1

Demo: Shuffled probe

# Control tasks

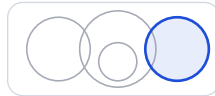




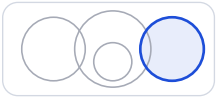

Hewitt and Liang (2019), Table 1

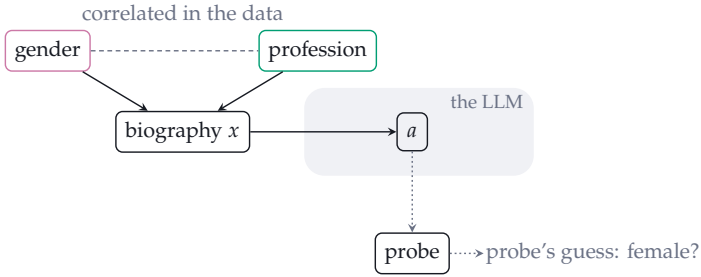
Demo: Shuffled probe

# Spurious correlations

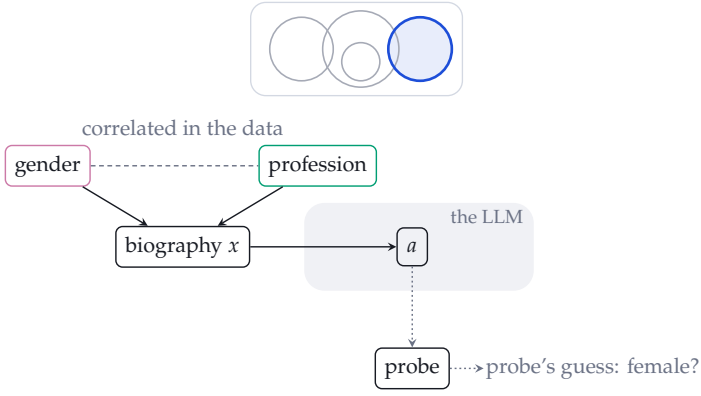


# Spurious correlations





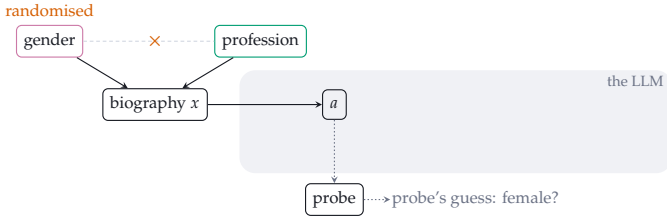
# Spurious correlations




In Bias in Bios, profession correlates strongly with gender.

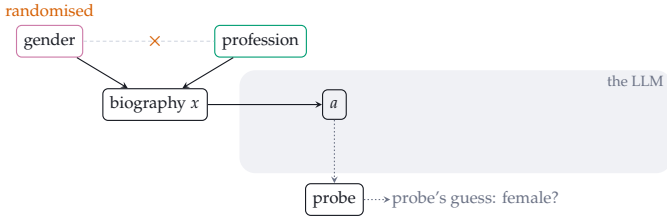
De-Arteaga et al. (2019)

# Randomization




Randomization can help us cut the link between profession and gender.

# Randomization

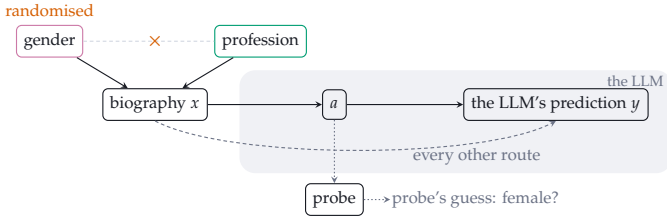



Randomization can help us cut the link between profession and gender.

Better still: we could change the biography to swap gender, but remain otherwise identical.

$x$ She has worked in hospitals for ten years. Mary . . .
$x'$ He has worked in hospitals for ten years. John . . .

# Randomization



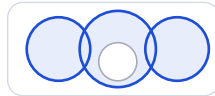

Randomization can help us cut the link between profession and gender.

Better still: we could change the biography to swap gender, but remain otherwise identical.

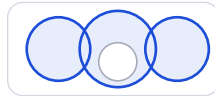
$x$ She has worked in hospitals for ten years. Mary ...
$x'$ He has worked in hospitals for ten years. John ...

Now we can measure the effect of gender on $a$, and on $y$.

# *What makes for a good probing experiment?*

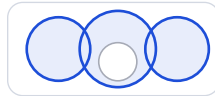


# *What makes for a good probing experiment?*



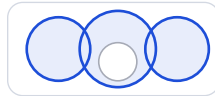
| | |
|---|---|
| Vary | labels (real or random) and inputs (real and with target concept changed) |

# *What makes for a good probing experiment?*



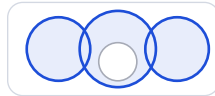
| | |
|---|---|
| Vary | labels (real or random) and inputs (real and with target concept changed) |
| Measure | selectivity, and how often the probe's label changes when the concept is modified |

# *What makes for a good probing experiment?*



| | |
|---|---|
| Vary | labels (real or random) and inputs (real and with target concept changed) |
| Measure | selectivity, and how often the probe's label changes when the concept is modified |
| Refuted if | selectivity is near zero, or the label rarely changes when the target concept is varied |

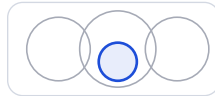
# *What makes for a good probing experiment?*



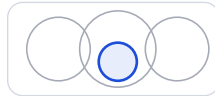
| | |
|---|---|
| Vary | labels (real or random) and inputs (real and with target concept changed) |
| Measure | selectivity, and how often the probe's label changes when the concept is modified |
| Refuted if | selectivity is near zero, or the label rarely changes when the target concept is varied |

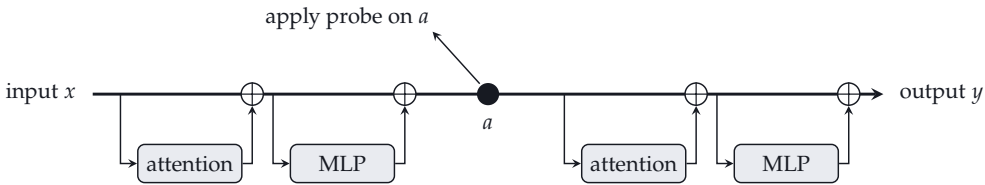
> Even still, a probe cannot show whether the model actually uses the target concept.

# Causal validation

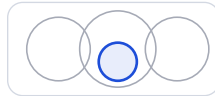


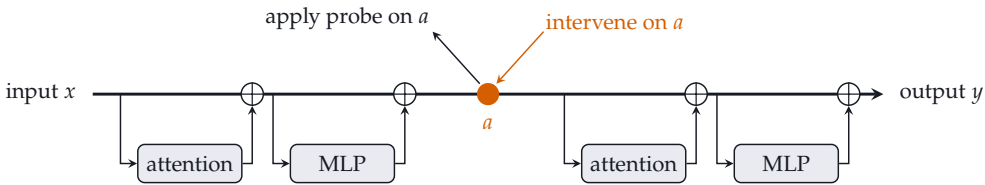
# Causal validation



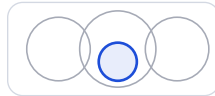


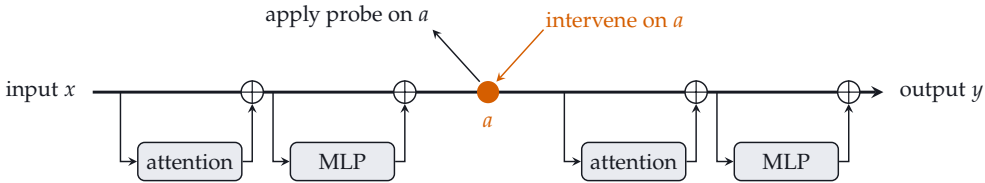
# Causal validation





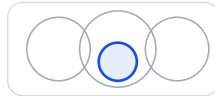
# Causal validation

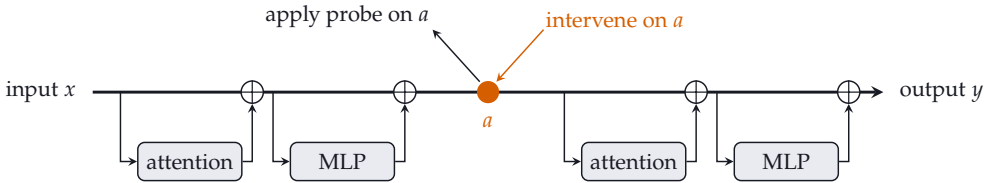





We observe $p(y \mid a)$

# Causal validation

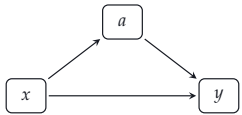





We observe $p(y \mid a)$

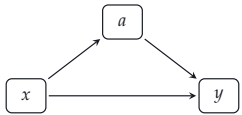
Intervene on $a$: $p(y \mid \mathrm{do}(a := \tilde{a}))$

# Counterfactual interventions




Pearl (2009); Joshi et al. (2026), Appendix A.1.1 and Table 3

# Counterfactual interventions





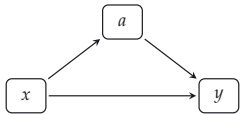


set $a$: $\tilde{y} = y(x, \text{do}(a := \tilde{a}))$

We hold the input fixed, but intervene on $a$ by setting it to some counterfactual value $\tilde{a}$.

Pearl (2009); Joshi et al. (2026), Appendix A.1.1 and Table 3

# Counterfactual interventions







set $a$: $\tilde{y} = y(x, \mathrm{do}(a := \tilde{a}))$

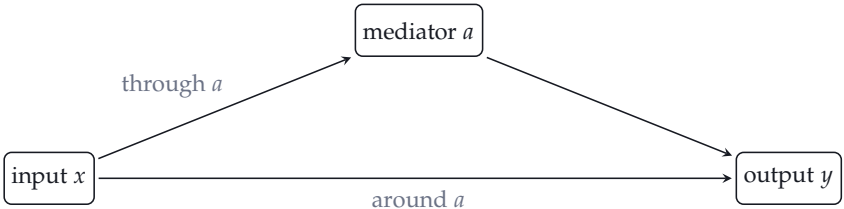
We hold the input fixed, but intervene on $a$ by setting it to some counterfactual value $\tilde{a}$.

**Intervention** Setting a variable to a value we choose, written $\mathrm{do}(a := \tilde{a})$, while every other part of the computation follows its usual rules.

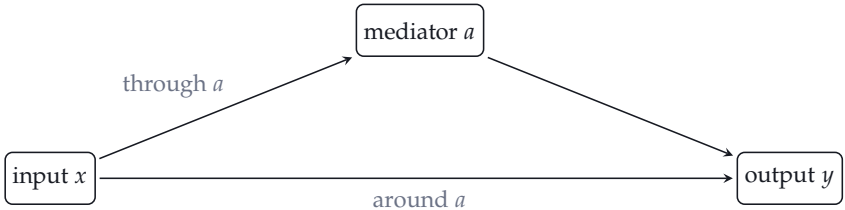
Randomising gender was an intervention too, on the input.

Pearl (2009); Joshi et al. (2026), Appendix A.1.1 and Table 3

# Causal mediation analysis

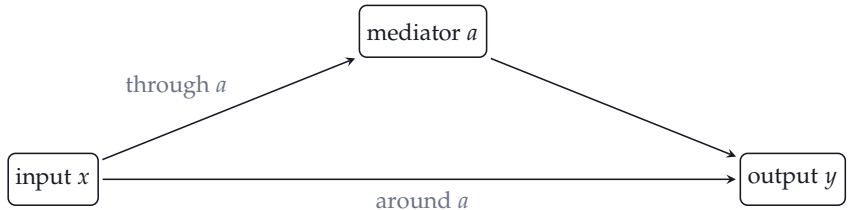


# Causal mediation analysis



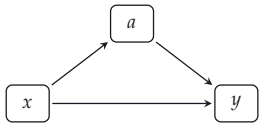

**Mediator** A part of the LLM on a route from the input to the output, such as $a$.

# Causal mediation analysis



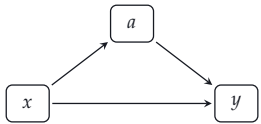

| | |
|---|---|
| Mediator | A part of the LLM on a route from the input to the output, such as $a$. |
| Mediation analysis | Changing the input and asking how much of the change in the output is from the mediator. |

# Indirect effects

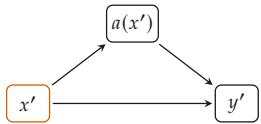



a biography

# Indirect effects



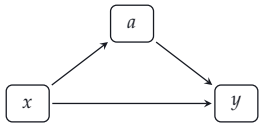

a biography



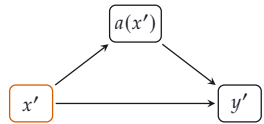

biography w/ gender swapped

change the input: total effect
$m(y') - m(y)$

# Indirect effects

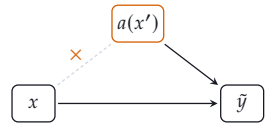



a biography




biography w/ gender swapped

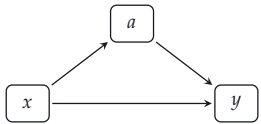
change the input: total effect $m(y') - m(y)$

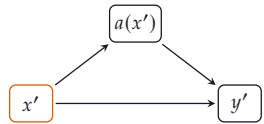



a biography

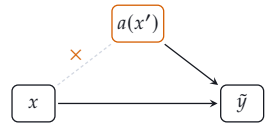
change only the mediator: indirect effect $m(\tilde{y}) - m(y)$

# Indirect effects



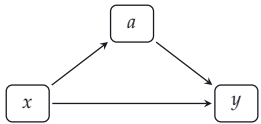

change the input: total effect $m(y') - m(y)$

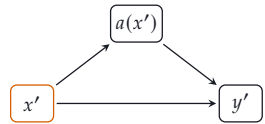



change only the mediator: indirect effect $m(\tilde{y}) - m(y)$

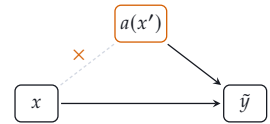
$$\text{IE} = m(\tilde{y}) - m(y), \text{ where } \tilde{y} = y(x, \text{do}(a := a(x')))$$

# Indirect effects




change the input: total effect $m(y') - m(y)$



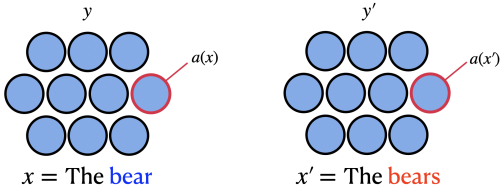

change only the mediator: indirect effect $m(\tilde{y}) - m(y)$

$$\text{IE} = m(\tilde{y}) - m(y), \text{ where } \tilde{y} = y(x, \text{do}(a := a(x')))$$

**Metric $m$** The target measurement (a scalar). Here, $m$ is the probability of "nurse", the model's output token given $x$.

# Computing indirect effects

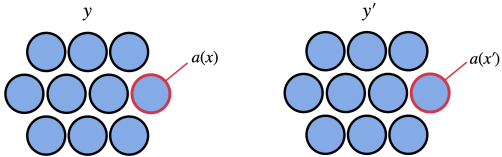
$$m = \log p(\text{sees}) - \log p(\text{see})$$




| Run | Input | Neuron $a$ | $m$ |
|---|---|---|---|
| 1 | $x$ | $a(x)$ | 2 |
| 2 | $x'$ | $a(x')$ | 6 |

# Computing indirect effects

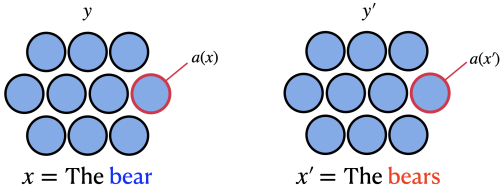
$$m = \log p(\text{sees}) - \log p(\text{see})$$




| Run | Input | Neuron $a$ | $m$ |
|---|---|---|---|
| 1 | $x$ | $a(x)$ | 2 |
| 2 | $x'$ | $a(x')$ | 6 |
| 3 | $x$ | $a(x')$ | 5 |
| 4 | $x'$ | $a(x)$ | 4 |

# Computing indirect effects

$$m = \log p(\text{sees}) - \log p(\text{see})$$

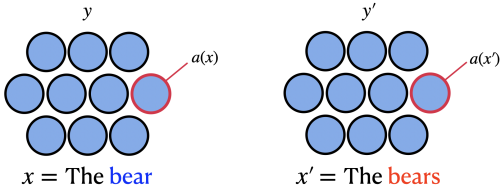



| Run | Input | Neuron $a$ | $m$ |
|---|---|---|---|
| 1 | $x$ | $a(x)$ | 2 |
| 2 | $x'$ | $a(x')$ | 6 |
| 3 | $x$ | $a(x')$ | 5 |
| 4 | $x'$ | $a(x)$ | 4 |

What is the **total effect**?

# Computing indirect effects

$$m = \log p(\text{sees}) - \log p(\text{see})$$

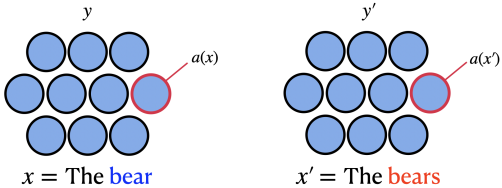



| Run | Input | Neuron $a$ | $m$ |
|---|---|---|---|
| 1 | $x$ | $a(x)$ | 2 |
| 2 | $x'$ | $a(x')$ | 6 |
| 3 | $x$ | $a(x')$ | 5 |
| 4 | $x'$ | $a(x)$ | 4 |

What is the **total effect**? $\text{TE} = m(y') - m(y) = 6 - 2 = 4$

# Computing indirect effects

$$m = \log p(\text{sees}) - \log p(\text{see})$$



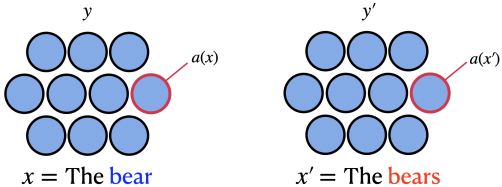

| Run | Input | Neuron $a$ | $m$ |
|---|---|---|---|
| 1 | $x$ | $a(x)$ | 2 |
| 2 | $x'$ | $a(x')$ | 6 |
| 3 | $x$ | $a(x')$ | 5 |
| 4 | $x'$ | $a(x)$ | 4 |

What is the **total effect**? $\text{TE} = m(y') - m(y) = 6 - 2 = 4$

What is the **indirect effect** of setting $a$ to $a'$ given $x$?

# Computing indirect effects

$$m = \log p(\text{sees}) - \log p(\text{see})$$



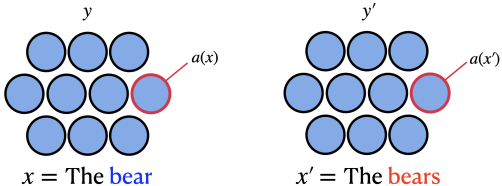

| Run | Input | Neuron $a$ | $m$ |
|---|---|---|---|
| 1 | $x$ | $a(x)$ | 2 |
| 2 | $x'$ | $a(x')$ | 6 |
| 3 | $x$ | $a(x')$ | 5 |
| 4 | $x'$ | $a(x)$ | 4 |

What is the **total effect**? TE $= m(y') - m(y) = 6 - 2 = 4$

What is the **indirect effect** of setting $a$ to $a'$ given $x$? IE $= m(x, \tilde{y}) - m(x, y) = 5 - 2 = 3$

# Computing indirect effects

$m = \log p(\text{sees}) - \log p(\text{see})$




| Run | Input | Neuron $a$ | $m$ |
|---|---|---|---|
| 1 | $x$ | $a(x)$ | 2 |
| 2 | $x'$ | $a(x')$ | 6 |
| 3 | $x$ | $a(x')$ | 5 |
| 4 | $x'$ | $a(x)$ | 4 |

What is the **total effect**? TE $= m(y') - m(y) = 6 - 2 = 4$

What is the **indirect effect** of setting $a$ to $a'$ given $x$? IE $= m(x, \tilde{y}) - m(x, y) = 5 - 2 = 3$

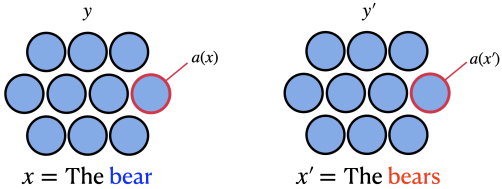
What if we patch in the reverse direction?

Tutorial worksheet, Problem 1

Parts 1 2 3

# Computing indirect effects

$m = \log p(\text{sees}) - \log p(\text{see})$




| Run | Input | Neuron $a$ | $m$ |
|---|---|---|---|
| 1 | $x$ | $a(x)$ | 2 |
| 2 | $x'$ | $a(x')$ | 6 |
| 3 | $x$ | $a(x')$ | 5 |
| 4 | $x'$ | $a(x)$ | 4 |

What is the **total effect**? TE $= m(y') - m(y) = 6 - 2 = 4$

What is the **indirect effect** of setting $a$ to $a'$ given $x$? IE $= m(x, \tilde{y}) - m(x, y) = 5 - 2 = 3$

What if we patch in the reverse direction? IE $= m(x', \tilde{y}) - m(x', y) = 4 - 6 = -2$

Tutorial worksheet, Problem 1

# *What does* IE/TE *measure?*

Claim

"This neuron holds 75% of the model's number representation, since IE/TE = 3/4."

Is this an accurate conclusion given the evidence?



# *What does* IE/TE *measure?*

Claim — "This neuron holds 75% of the model's number representation, since IE/TE = 3/4."

Is this an accurate conclusion given the evidence?

Supported — On this pair, for this metric and this direction, setting $a$ alone moves $m$ by 75% of the total change.



# *What does* IE/TE *measure?*

Claim — "This neuron holds 75% of the model's number representation, since IE/TE = 3/4."

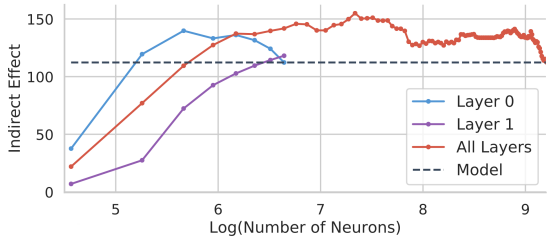
Is this an accurate conclusion given the evidence?

Supported — On this pair, for this metric and this direction, setting *a* alone moves *m* by 75% of the total change.

> The ratio is a share of a *change in the output*.



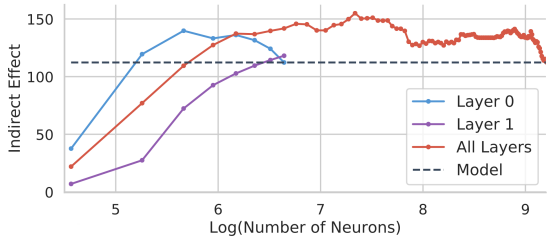
# Indirect effects of sets of neurons




If we rank neurons by their indirect effects, greedily add them to a set, and intervene on the set, we see the trend above as we increase the number of neurons in the set. What does the decline near the right of the graph tell us?

Vig et al. (2020)

# Indirect effects of sets of neurons




If we rank neurons by their indirect effects, greedily add them to a set, and intervene on the set, we see the trend above as we increase the number of neurons in the set. What does the decline near the right of the graph tell us?

That some neurons have *negative* indirect effect, and thus counteract the effect of the other neurons.

Vig et al. (2020)

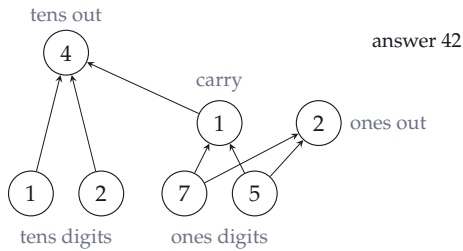
# Activation patching

$$\begin{array}{rcl} & \color{#3b82f6}{1} & \\ & 1 & 7 \\ + & 2 & 5 \\ \hline & 4 & 2 \end{array}$$

# Activation patching

$$\begin{array}{r} {\scriptstyle 1}\phantom{0} \\ 1\ 7 \\ +\ 2\ 5 \\ \hline 4\ 2 \end{array}$$

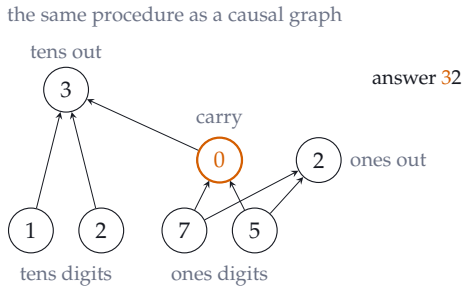
the same procedure as a causal graph




Mueller et al. (2025)

# Activation patching

$$\begin{array}{r} {\scriptstyle 1}\phantom{00} \\ 1\ 7 \\ +\ 2\ 5 \\ \hline 4\ 2 \end{array}$$

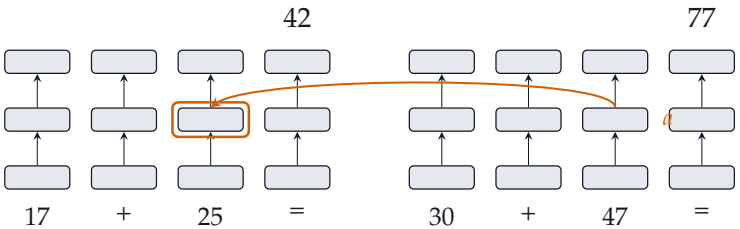



This is an intervention on the program: do(carry := 0).

# Activation patching

$$\begin{array}{rrr} & 3 & 0 \\ + & 4 & 7 \\ \hline & 7 & 7 \end{array}$$

# Activation patching

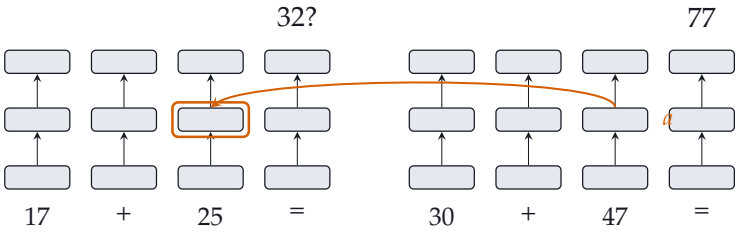



$x = 17 + 25$, the run we patch

$x' = 30 + 47$, the run we copy from

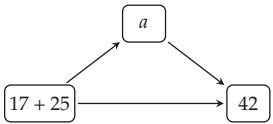
$$\begin{array}{rr} & 3\ 0 \\ + & 4\ 7 \\ \hline & 7\ 7 \end{array}$$

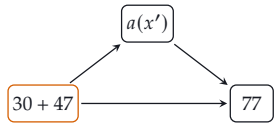
# Activation patching



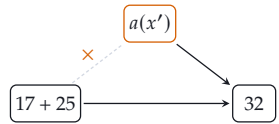

If $a$ encodes the carry operation, then the network should now answer 32.

# Activation patching



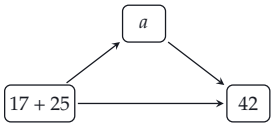



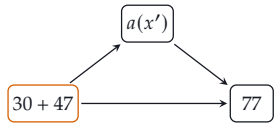

change the input




change only the mediator

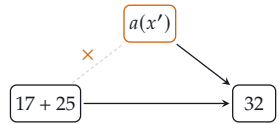
# Activation patching




change the input

total effect: 42 → 77

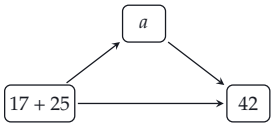


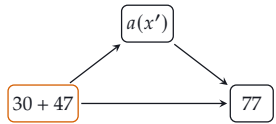

change only the mediator

indirect effect through $a$: 42 → 32

Patching is the third panel, run on real inputs.

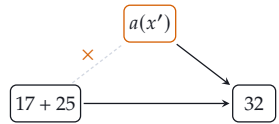
# Activation patching




change the input

total effect: 42 → 77




change only the mediator

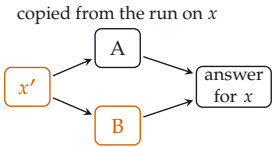
indirect effect through $a$: 42 → 32

Patching is the third panel, run on real inputs.

On the biographies, we copy $a$ from the he-run into the she-run.

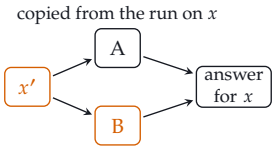
# Necessity and sufficiency

Is A sufficient?

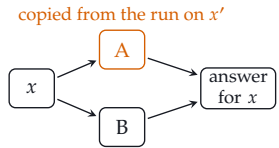


# Necessity and sufficiency

Is A sufficient?

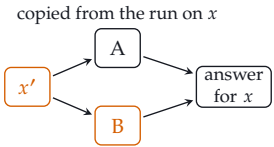



Is A necessary?

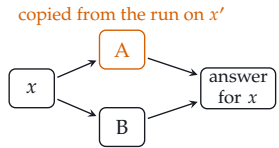


# Necessity and sufficiency

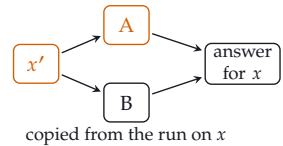
Is A sufficient?




Is A necessary?

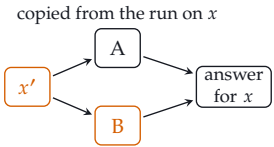



Is B sufficient?

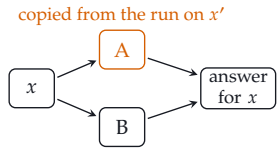


# Necessity and sufficiency

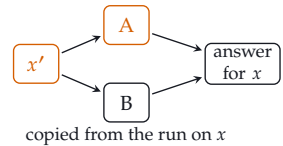
Is A sufficient?




Is A necessary?



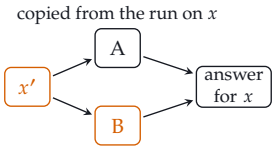

Is B sufficient?



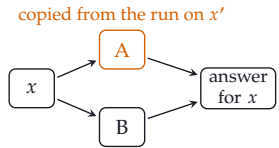

**Sufficient** Copying this part alone from the run on $x$ brings back $x'$s answer.

# Necessity and sufficiency

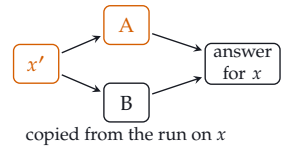
Is A sufficient?




Is A necessary?

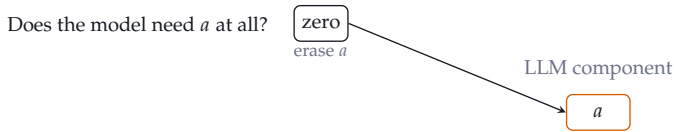



Is B sufficient?

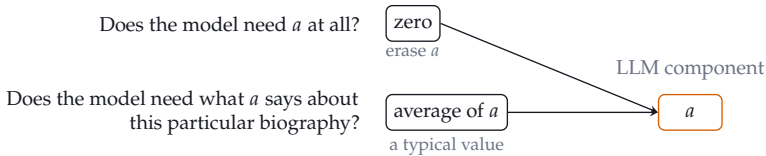



| | |
|---|---|
| Sufficient | Copying this part alone from the run on $x$ brings back $x'$'s answer. |
| Necessary | Replacing this part alone with its value from $x'$ changes the answer. |

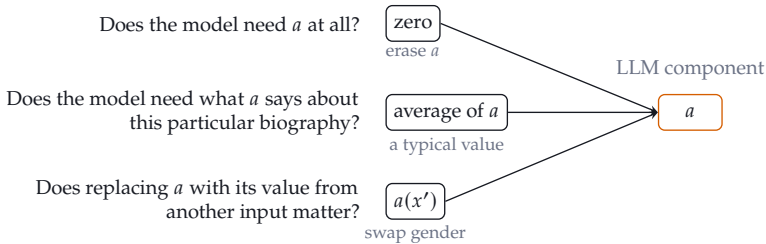
# *How should we patch a?*



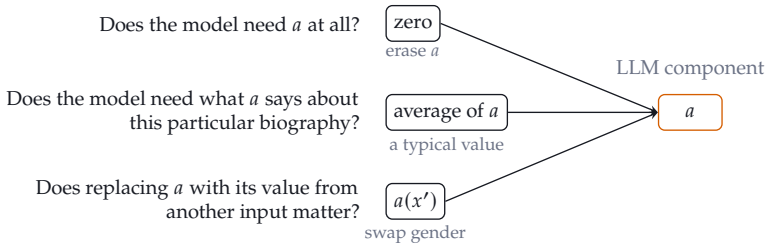
# *How should we patch a?*



# *How should we patch $a$?*

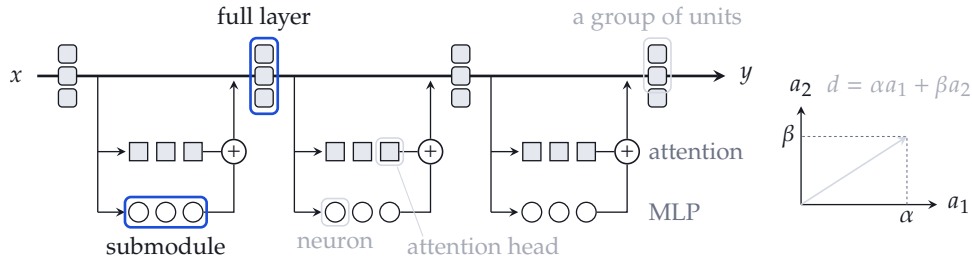


# *How should we patch a?*



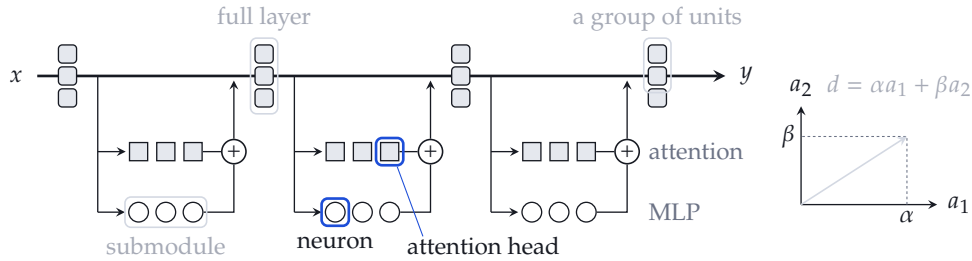

**Intervention class** The kinds of change an experiment allows, including the value we put in.

# Causal mediators in LLMs




What kind of component(s) should we use as mediators?

# Causal mediators in LLMs

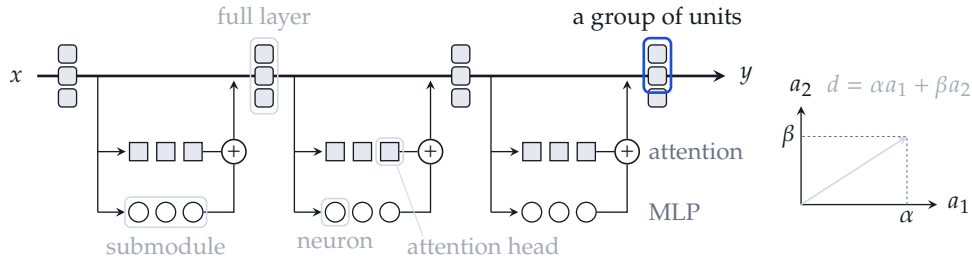



What kind of component(s) should we use as mediators?

**Neurons and heads**: fine-grained and tractable to search over, but they often respond to many seemingly unrelated concepts.

Mueller et al. (2026); Bricken et al. (2023)

# Causal mediators in LLMs

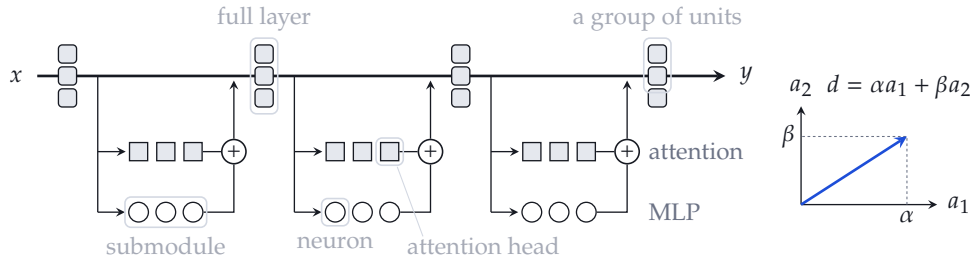



What kind of component(s) should we use as mediators?

**Groups of neurons/heads**: combinatorial explosion. :(

Mueller et al. (2026); Bricken et al. (2023)

# Causal mediators in LLMs

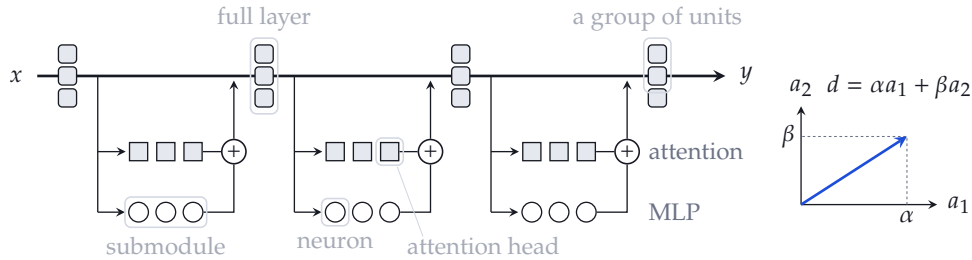



What kind of component(s) should we use as mediators?

**Directions**: often interpretable, but must be learned using some external component like a probe or an SAE.

Mueller et al. (2026); Bricken et al. (2023)

# Causal mediators in LLMs



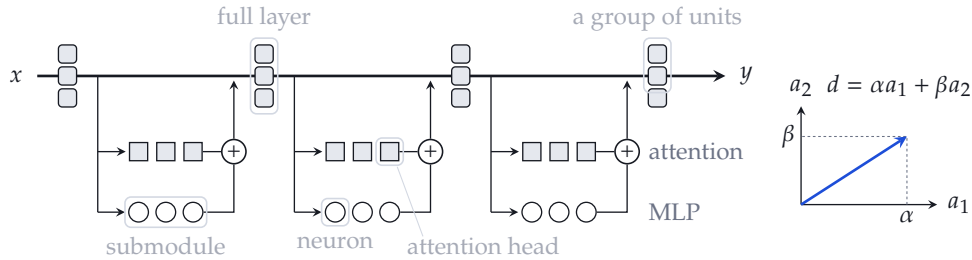

What kind of component(s) should we use as mediators?

**Directions**: often interpretable, but must be learned using some external component like a probe or an SAE.

A featurizer $\varphi$ can be learned to map activations to such directions.

Mueller et al. (2026); Bricken et al. (2023)

# Causal mediators in LLMs



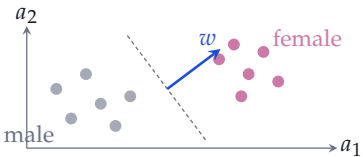

What kind of component(s) should we use as mediators?

**Directions**: often interpretable, but must be learned using some external component like a probe or an SAE.

A featurizer $\varphi$ can be learned to map activations to such directions.

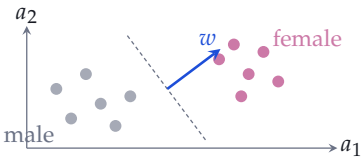
Mueller et al. (2026); Bricken et al. (2023)

# Featurization

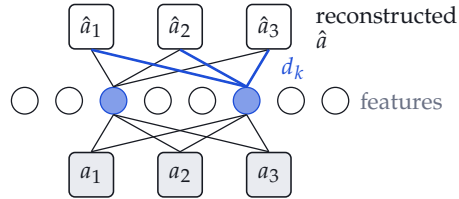



Probe: trained to classify with labeled data

# Featurization

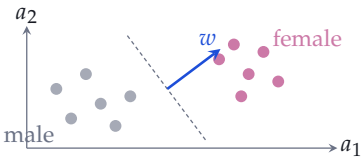



Probe: trained to classify with labeled data

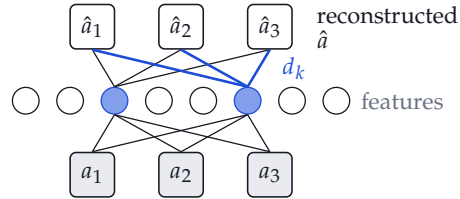



Sparse autoencoder (SAE): trained to reconstruct $a$ without labels.
Yields multiple directions

# Featurization




Probe: trained to classify with labeled data

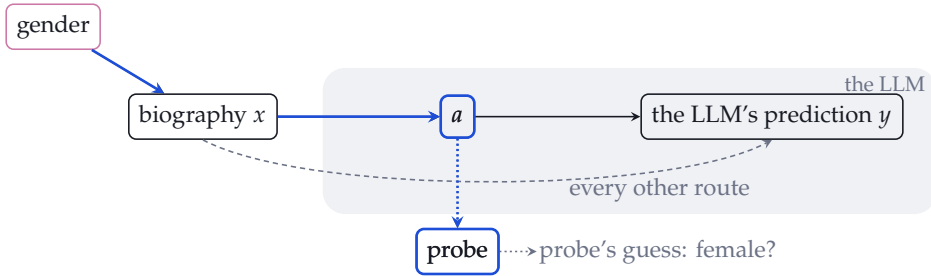



Sparse autoencoder (SAE): trained to reconstruct $a$ without labels.
Yields multiple directions

| | |
|---|---|
| Sparse autoencoder (SAE) | A small network trained, without labels, to reconstruct an activation from a few of many learned directions. |

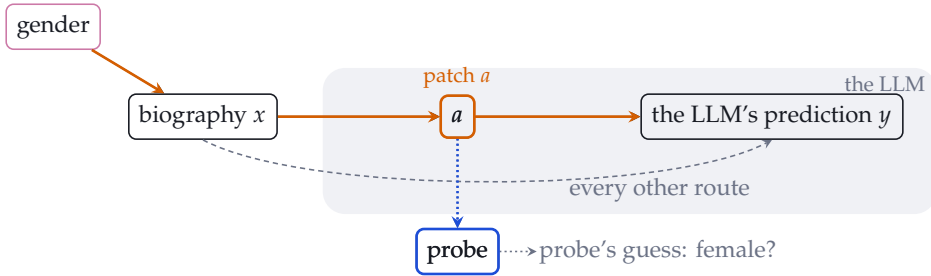
Mueller et al. (2026)

# Probing vs. patching




Probing asks: can information about gender be decoded from $a$?

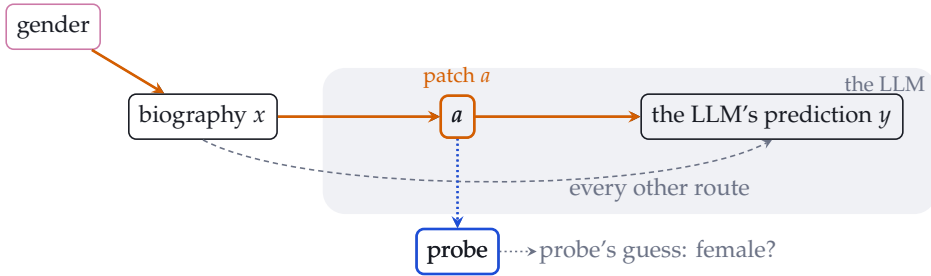
# Probing vs. patching




Probing asks: can information about gender be decoded from $a$?

Patching asks: does changing $a$ change the LLM's prediction?

# Probing vs. patching

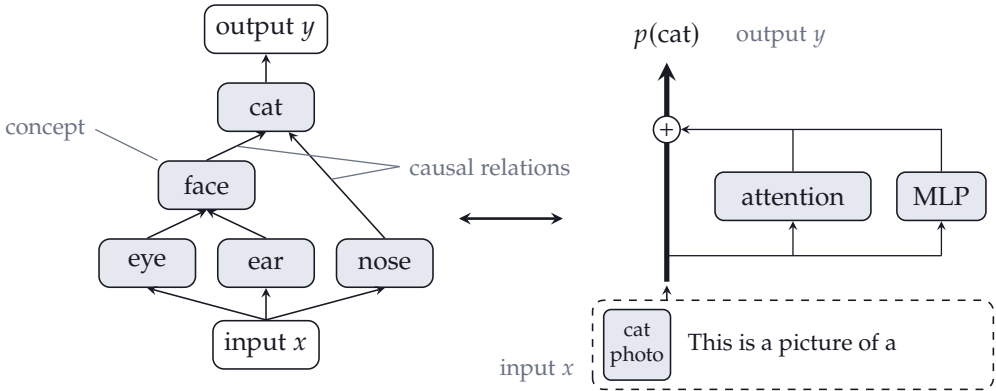



Probing asks: can information about gender be decoded from $a$?

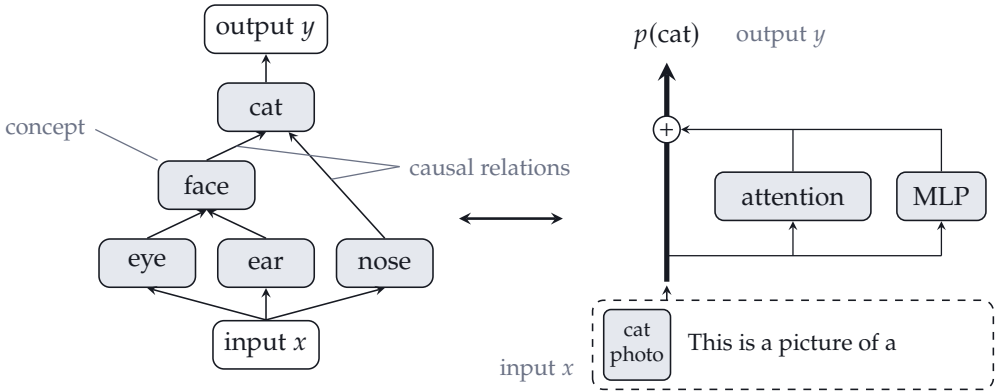
Patching asks: does changing $a$ change the LLM's prediction?

For example, if we claim that an LLM relies on gender, that would be a claim relating to the second question.

# Mechanisms


output $y$
cat
concept
face
causal relations
eye
ear
nose
input $x$
$p$(cat)
output $y$
attention
MLP
cat photo
This is a picture of a
input $x$

# Mechanisms




What if we have no names for these concepts?

# Summary

Probing

*a* correlates with gender.

# Summary

Probing

$a$ correlates with gender.

Patching

$a$ influences the model's use of gendered terms.

# Summary

**Probing**

$a$ correlates with gender.

**Patching**

$a$ influences the model's use of gendered terms.

**Next**

What exactly can we conclude from these results?

# Pointers to the literature

## Probing

Hewitt and Liang (2019): *Designing and Interpreting Probes with Control Tasks*

Hewitt and Manning (2019): *A Structural Probe for Finding Syntax in Word Representations*

Voita and Titov (2020): *Information-Theoretic Probing with Minimum Description Length*

Ravichander et al. (2021): *Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance?*

Belinkov (2022): *Probing Classifiers: Promises, Shortcomings, and Advances*

# Pointers to the literature

## Probing with interventions

Ravfogel et al. (2020): *Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection*

Elazar et al. (2021): *Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals*

Belrose et al. (2023): *LEACE: Perfect linear concept erasure in closed form*

# Pointers to the literature

## Paired inputs and gender

De-Arteaga et al. (2019): *Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting*

Kaushik et al. (2020): *Learning the Difference that Makes a Difference with Counterfactually-Augmented Data*

Vig et al. (2020): *Investigating Gender Bias in Language Models Using Causal Mediation Analysis*

# Pointers to the literature

## Causal mediation and activation patching

Robins and Greenland (1992): *Identifiability and Exchangeability for Direct and Indirect Effects*

Pearl (2001): *Direct and Indirect Effects*

Meng et al. (2022): *Locating and Editing Factual Associations in GPT*

Wang et al. (2023): *Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small*

Heimersheim and Nanda (2024): *How to use and interpret activation patching*

Zhang and Nanda (2024): *Towards Best Practices of Activation Patching in Language Models: Metrics and Methods*

Mueller et al. (2026): *The Quest for the Right Mediator: Surveying Mechanistic Interpretability for NLP Through the Lens of Causal Mediation Analysis*

The full bibliography is on the tutorial website.

# Outline

Asking causal questions — Aaron Mueller
causal mediation

Levels of causal questions
the causal ladder

Identifying the answers
when are answers to causal questions unique?

# Outline

Asking causal questions

causal mediation

Levels of causal questions Patrik Reizinger

the causal ladder

Identifying the answers

when are answers to causal questions unique?

# *Why do we need causality?*

# Probe example

the claim, and its evidence

"The model represents gender at this layer, and the prediction relies on it"

a linear probe on these bios

Claim from demo 11; decoding versus control after Joshi et al. (2026), Appendix D.2

Demo: Claim diagnosis

# Probe example

| the claim, and its evidence | a reader can take it to say |
| --- | --- |
| “The model represents gender at this layer, and the prediction relies on it”<br>a linear probe on these bios | the network uses gender here for its prediction |

Claim from demo 11; decoding versus control after Joshi et al. (2026), Appendix D.2

Demo: Claim diagnosis

# Probe example

| the claim, and its evidence | a reader can take it to say | the evidence supports |
|---|---|---|
| "The model represents gender at this layer, and the prediction relies on it"<br>a linear probe on these bios | the network uses gender here for its prediction | gender is decodable from $a$ on these bios |

Claim from demo 11; decoding versus control after Joshi et al. (2026), Appendix D.2

Demo: Claim diagnosis

# Probe example

| the claim, and its evidence | a reader can take it to say | the evidence supports |
|---|---|---|
| "The model represents gender at this layer, and the prediction relies on it"<br>a linear probe on these bios | the network uses gender here for its prediction | gender is decodable from $a$ on these bios |

decoding, what the probe did: read $a$




control, what "relies on it" claims: set $a$, does $y$ move?




Claim from demo 11; decoding versus control after Joshi et al. (2026), Appendix D.2

Demo: Claim diagnosis

# Patching head 9.6 changes the IOI answer, but there might be another route

the claim, and its evidence

A name-mover head (9.6 in GPT-2 small) contributes to the IOI behaviour

path patching

Claim from demo 11, paraphrasing Wang et al. (2023); the second route after Joshi et al. (2026), Appendix D.2

# Patching head 9.6 changes the IOI answer, but there might be another route

| the claim, and its evidence | a reader can take it to say |
| --- | --- |
| A name-mover head (9.6 in GPT-2 small) contributes to the IOI behaviour<br>path patching | head 9.6 plays a part in the IOI answer, or it is *the* IOI mechanism |

Claim from demo 11, paraphrasing Wang et al. (2023); the second route after Joshi et al. (2026), Appendix D.2

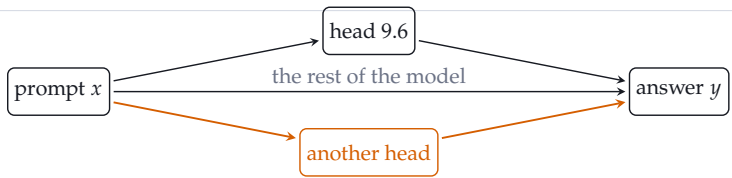
# Patching head 9.6 changes the IOI answer, but there might be another route

| the claim, and its evidence | a reader can take it to say | the evidence supports |
|---|---|---|
| A name-mover head (9.6 in GPT-2 small) contributes to the IOI behaviour<br>path patching | head 9.6 plays a part in the IOI answer, or it is *the* IOI mechanism | patching what head 9.6 sends changes the tested IOI answers |

Claim from demo 11, paraphrasing Wang et al. (2023); the second route after Joshi et al. (2026), Appendix D.2

# Patching head 9.6 changes the IOI answer, but there might be another route

| the claim, and its evidence | a reader can take it to say | the evidence supports |
|---|---|---|
| A name-mover head (9.6 in GPT-2 small) contributes to the IOI behaviour<br>path patching | head 9.6 plays a part in the IOI answer, or it is *the* IOI mechanism | patching what head 9.6 sends changes the tested IOI answers |




the patch shows that the route through 9.6 changes the answer; another route can give the same answer, so 9.6 is one mechanism, not *the* mechanism.

Claim from demo 11, paraphrasing Wang et al. (2023); the second route after Joshi et al. (2026), Appendix D.2

# The carry examples: claim ambiguity

the claim, and its evidence

"This activation encodes the carry"

activation patching over the tested sums

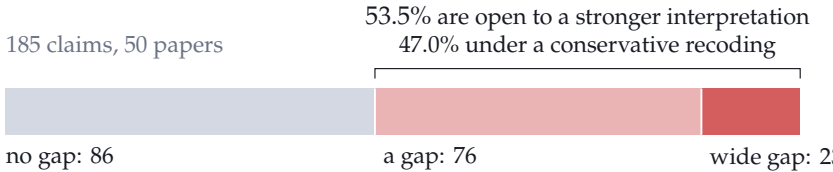
# The carry examples: claim ambiguity

| the claim, and its evidence | a reader can take it to say |
|---|---|
| "This activation encodes the carry"<br>activation patching over the tested sums | the carry is decodable from it, or it is the carry the network uses |

# The carry examples: claim ambiguity

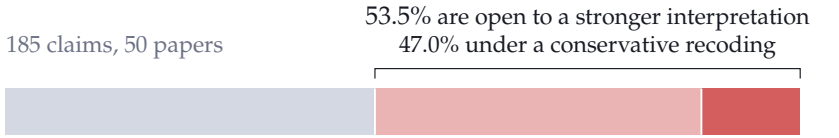
| the claim, and its evidence | a reader can take it to say | the evidence supports |
|---|---|---|
| "This activation encodes the carry"<br>activation patching over the tested sums | the carry is decodable from it, or it is the carry the network uses | copying it from $30+47$ changes the answers of the tested sums that carry |

# Formulating precise (causal) claims is hard



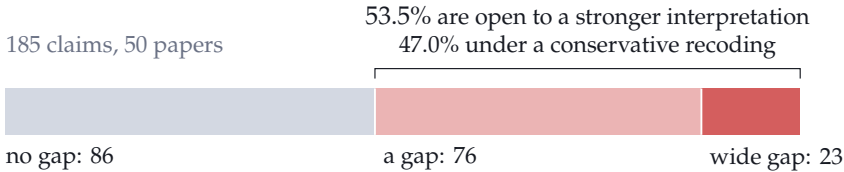

Joshi et al. (2026), Appendix G (pilot study)

# Formulating precise (causal) claims is hard




Joshi et al. (2026), Appendix G (pilot study)

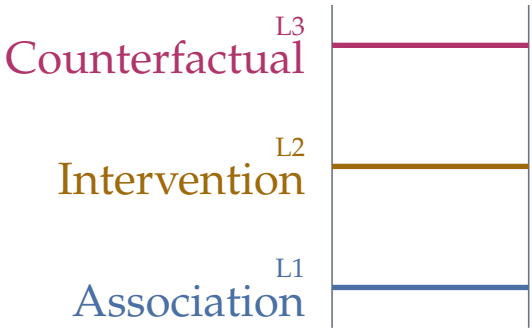
# Formulating precise (causal) claims is hard




The gap measures what a reader takes away, not what the authors meant: *the* circuit can be a name, and "encodes" often means "is decodable from".

Joshi et al. (2026), Appendix G (pilot study)

# Pearl's ladder sorts questions into rungs




$y_{a \leftarrow \tilde{a}}(x_0)$: what would this input have produced had $a$ been $\tilde{a}$?

$p(y \mid \mathrm{do}(a := \tilde{a}))$: does $y$ change when we set $a$?

$p(y \mid a)$: does $y$ go with $a$ across inputs?

Joshi et al. (2026), Appendix D.1

$x$ input · $a$ activation at layer $\ell$ · $y$ prediction · Demo: Causal ladder

# Mapping interpretability questions to rungs makes them better defined



**Association**

| | |
|---|---|
| $p(y \mid x)$ | which outputs are likely for this prompt? next-token prediction |
| $p(a \mid x)$ | which activations does this input produce? probing, SAE features |
| $p(y \mid a)$ | given these activations, which outputs are likely? linear probes |

Joshi et al. (2026), Appendix D.1

$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

# Mapping interpretability questions to rungs makes them better defined



**Intervention**

| | |
|---|---|
| $p(y \mid \mathrm{do}(x := x'))$ | if I change the prompt, how does the output change? minimal pairs |
| $p(y \mid \mathrm{do}(a := \tilde{a}))$ | if I patch $a$, how does the output change? patching, ablation |
| $p(y \mid \mathrm{do}(\text{feature}))$ | if I move a direction in $a$, what changes? steering, feature ablation |

**Association**

| | |
|---|---|
| $p(y \mid x)$ | which outputs are likely for this prompt? next-token prediction |
| $p(a \mid x)$ | which activations does this input produce? probing, SAE features |
| $p(y \mid a)$ | given these activations, which outputs are likely? linear probes |

Joshi et al. (2026), Appendix D.1

$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

# Mapping interpretability questions to rungs makes them better defined



**Counterfactual**

$y_{x \leftarrow x'}(x_0)$ — for this run, would the output have differed with prompt $x'$?

$y_{a \leftarrow \tilde{a}}(x_0)$ — would this generation have changed had $a$ been $\tilde{a}$? causal scrubbing

$y_{\text{feature} \leftarrow \cdot}(x_0)$ — would this behaviour have occurred without this feature? necessity tests

**Intervention**

$p(y \mid \text{do}(x := x'))$ — if I change the prompt, how does the output change? minimal pairs

$p(y \mid \text{do}(a := \tilde{a}))$ — if I patch $a$, how does the output change? patching, ablation

$p(y \mid \text{do}(\text{feature}))$ — if I move a direction in $a$, what changes? steering, feature ablation

**Association**

$p(y \mid x)$ — which outputs are likely for this prompt? next-token prediction

$p(a \mid x)$ — which activations does this input produce? probing, SAE features

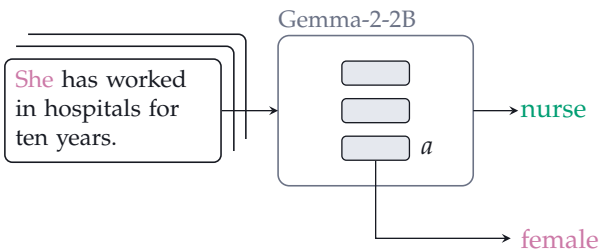
$p(y \mid a)$ — given these activations, which outputs are likely? linear probes

Joshi et al. (2026), Appendix D.1

$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

# Gender and profession: causal questions for each rung

*Do these activations correlate with gender?*





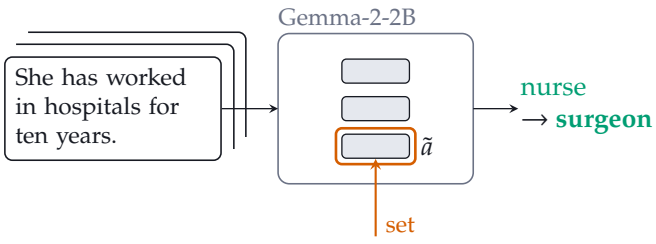

association (a probe): gender is decodable from this layer.

not supported: the model represents gender here.

Rungs after Pearl and Mackenzie (2018); rows adapted from Joshi et al. (2026); Belinkov (2022); Heimersheim and Nanda (2024)

# Gender and profession: causal questions for each rung

*Does setting these activations change the profession the model predicts?*





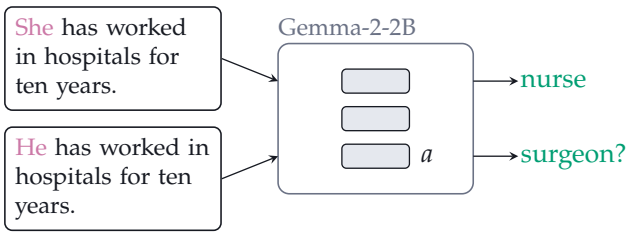

intervention (patching, ablation, steering): setting $a$ here changes the prediction, on the tested bios.

not supported: *the* gender mechanism.

Rungs after Pearl and Mackenzie (2018); rows adapted from Joshi et al. (2026); Belinkov (2022); Heimersheim and Nanda (2024)

# Gender and profession: causal questions for each rung

*Would the model have predicted surgeon?*





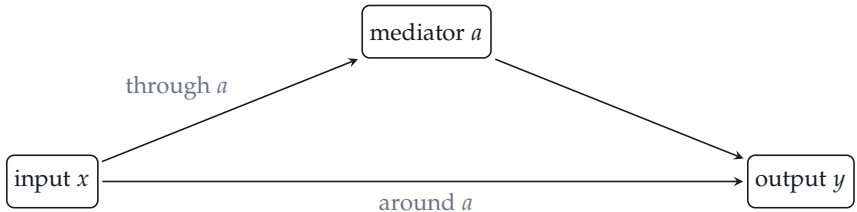

counterfactual (the swapped bio, rerun): if the swap changes gender and nothing else, this bio would have been predicted surgeon.

not supported: other bios.

Rungs after Pearl and Mackenzie (2018); rows adapted from Joshi et al. (2026); Belinkov (2022); Heimersheim and Nanda (2024)

*What is the difference between interventions and counterfactuals?*

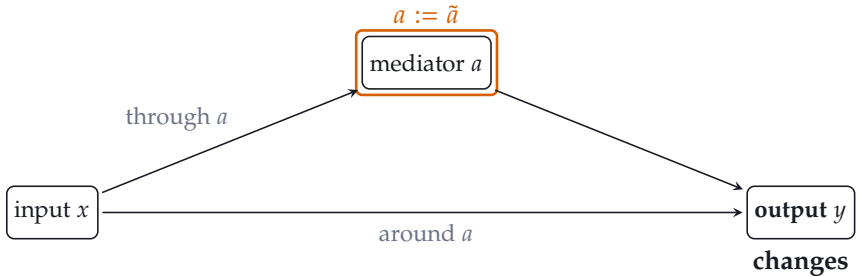
# Setting a mediator changes the output without showing how




Mediation as in Part 1; Pearl (2001)

$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

# Setting a mediator changes the output without showing how

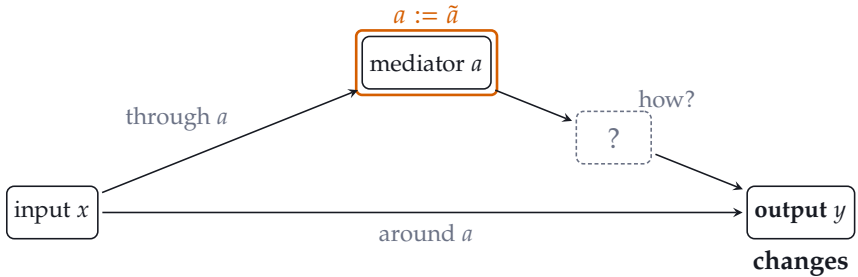



Setting $a$ is enough to change $y$.

Mediation as in Part 1; Pearl (2001)

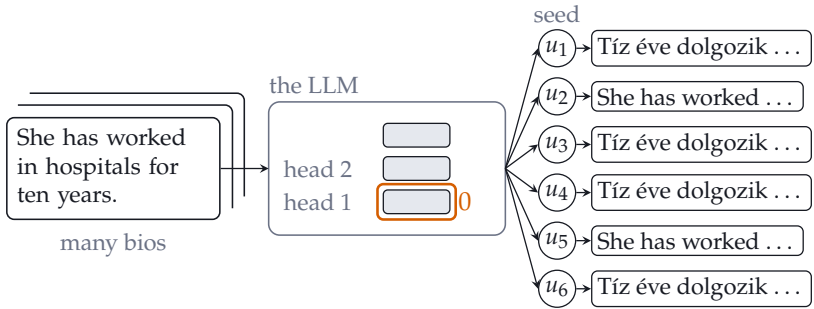
$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

# Setting a mediator changes the output without showing how




Setting $a$ is enough to change $y$, but not enough to understand the mechanism.

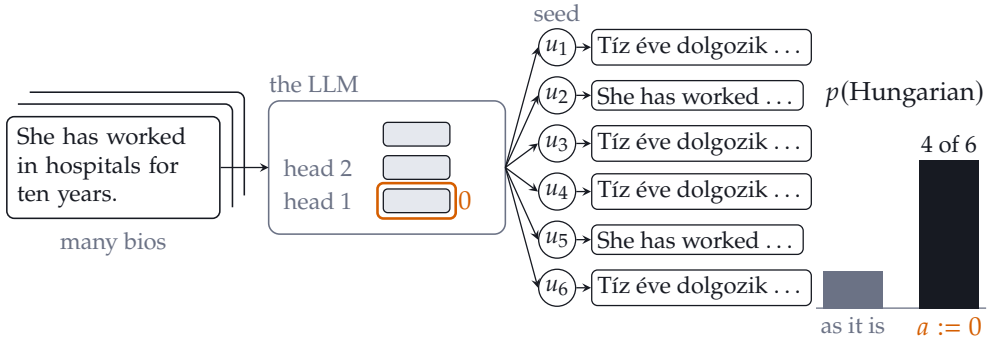
# Zeroing this head makes Hungarian more likely *on average*



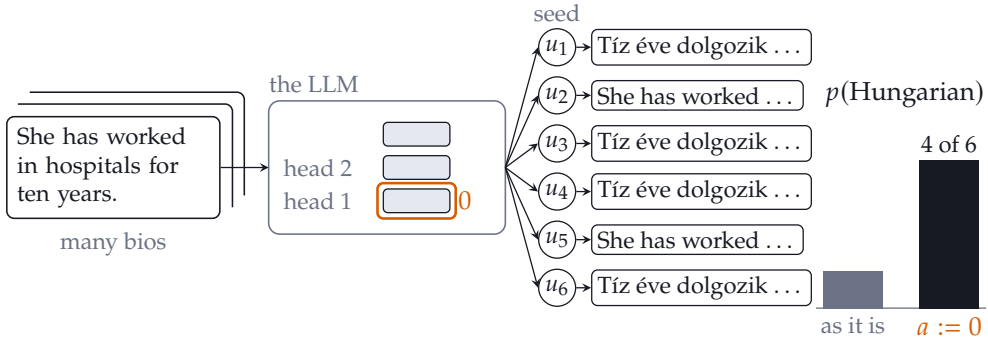

Schematic; do after Pearl (2009)

$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

# Zeroing this head makes Hungarian more likely *on average*



# Zeroing this head makes Hungarian more likely *on average*




$p(\text{Hungarian} \mid \text{do}(a := 0))$ is about the distribution: we are sampling.

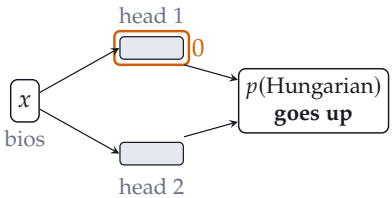
Schematic; do after Pearl (2009)

$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

# Intervening on a head does not show uniqueness

Part 1: a route can be sufficient without being necessary.

zero head 1: one way

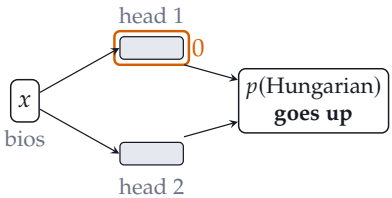



$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

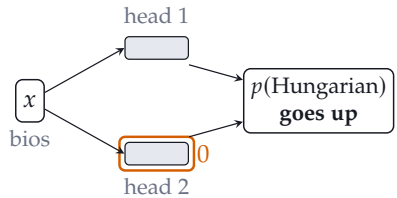
# Intervening on a head does not show uniqueness

Part 1: a route can be sufficient without being necessary.

zero head 1: one way




zero head 2: another way

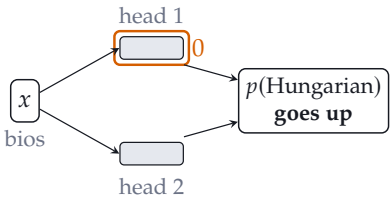



$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

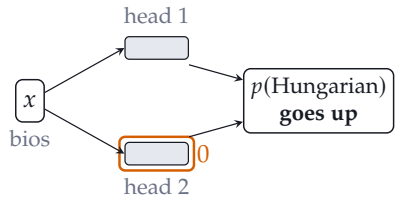
# Intervening on a head does not show uniqueness

Part 1: a route can be sufficient without being necessary.

zero head 1: one way



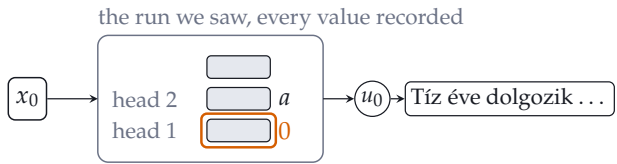

zero head 2: another way




Rung two is "local": it shows sufficiency, not necessity.

$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

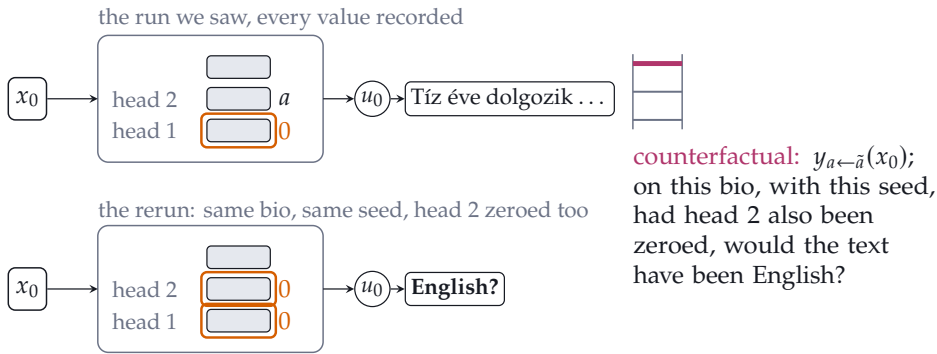
# A counterfactual is about a single sample in an imagined world




Huszár (2019); Pearl (2009)

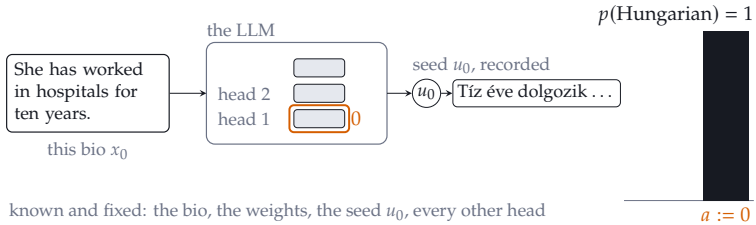
$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

# A counterfactual is about a single sample in an imagined world



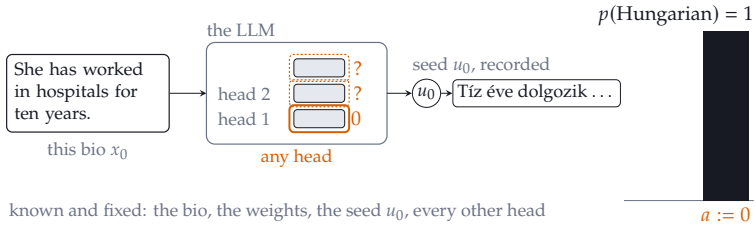

counterfactual: $y_{a \leftarrow \tilde{a}}(x_0)$; on this bio, with this seed, had head 2 also been zeroed, would the text have been English?

# We still zero the head, but now we know everything else




$p$(Hungarian) is 1 or 0: the inference is deterministic.

# We still zero the head, but now we know everything else



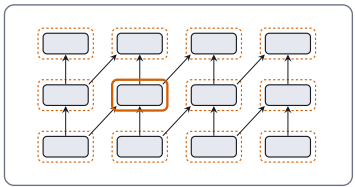

$p$(Hungarian) is 1 or 0: the inference is deterministic.

With everything else fixed and known, we can rerun this output with any head changed.

# *What does causality bring to MechInterp?*

# An open model gives us full control over the causal graph!

open model, one input $x_0$



every activation can be set, and the same input rerun

clinical trial, a population

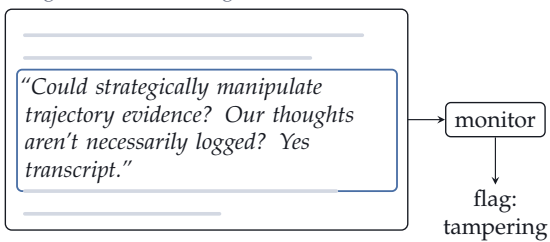



one variable randomised; the answer is an average over patients

# *How do we climb up the ladder?*

# A chain-of-thought monitor might only need rung one

*Do we always need a stronger claim?*






association: if lines like this go with tampering, they predict it.

not supported: what makes the agent tamper; once the line and the tampering come apart, the monitor misses.

Chain of thought from METR (2026), p. 17; the monitor is our illustration

# Controlling deception does not need counterfactuals



A distribution

If we only need how often the output is Hungarian across prompts, $p(\text{Hungarian} \mid \text{do}(a := 0))$ suffices.

$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

# Controlling deception does not need counterfactuals



A distribution

If we only need how often the output is Hungarian across prompts, $p(\text{Hungarian} \mid \text{do}(a := 0))$ suffices.



Control

Prohibiting deception is useful even without understanding its mechanism.

$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

# The type of data can give a higher rung

| the data | | the rung it reaches |
|---|---|---|
|   | i.i.d. bios as they come: we can only compute correlations (Part 1's probe) |  association |

Rungs after Pearl and Mackenzie (2018); De-Arteaga et al. (2019); Vig et al. (2020)

$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

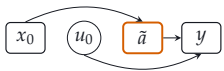
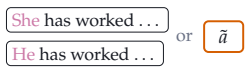
# The type of data can give a higher rung

| the data | | the rung it reaches |
|---|---|---|
| She has worked . . . / He has worked . . . or $\tilde{a}$ | data we control: we set $a$ (patching), or swap gender and keep the profession | intervention |
| She has . . . → probe | i.i.d. bios as they come: we can only compute correlations (Part 1's probe) | association |

Rungs after Pearl and Mackenzie (2018); De-Arteaga et al. (2019); Vig et al. (2020)

$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

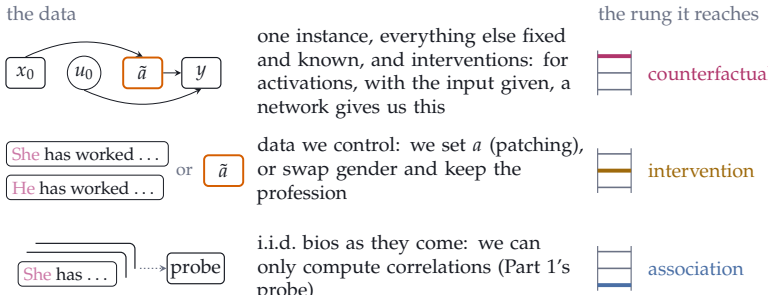
# The type of data can give a higher rung

| the data | | the rung it reaches |
|---|---|---|
|   | one instance, everything else fixed and known, and interventions: for activations, with the input given, a network gives us this |  counterfactual |
|   | data we control: we set $a$ (patching), or swap gender and keep the profession |  intervention |
|   | i.i.d. bios as they come: we can only compute correlations (Part 1's probe) |  association |

Rungs after Pearl and Mackenzie (2018); De-Arteaga et al. (2019); Vig et al. (2020)

$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

# The type of data can give a higher rung




With greedy decoding, rerunning one input gives the rung-three answer about activations; sampling multiple reruns gives rung two.

Rungs after Pearl and Mackenzie (2018); De-Arteaga et al. (2019); Vig et al. (2020)

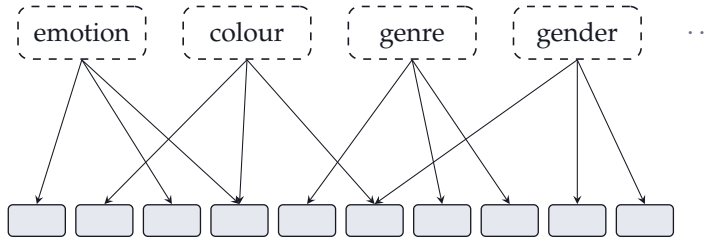
$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

# We interpret activations, but we care about concepts

Concepts are informal, high-level things; activations can represent them mathematically. What is the map between concepts and activations?

**concepts**
what claims name;
never observed,
not set directly

**activation $a$**
read and set in
every forward pass

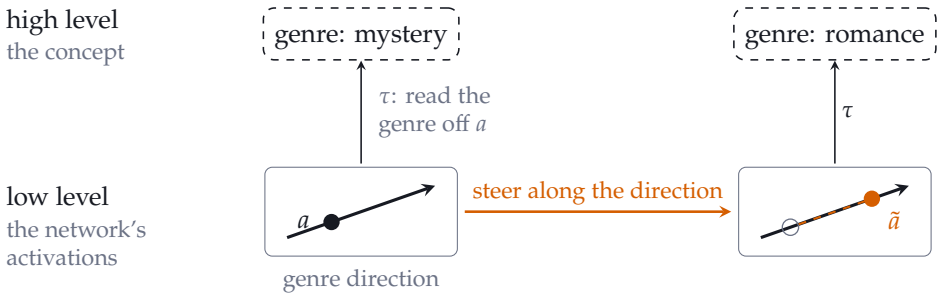



one layer, one position

$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

# If we want to change a concept, we need to find it in activation space

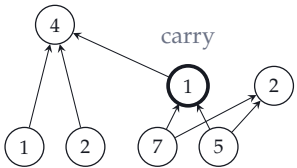
Concept: the genre. Representation: a direction in $a$.




Schematic; the direction is what distributed alignment search (DAS) learns, next frames

# Distributed Alignment Search learns an activation subspace that maps to a concept

1. select a concept



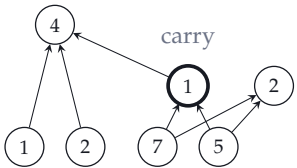

Schematic; Geiger et al. (2024); Sutter et al. (2025)

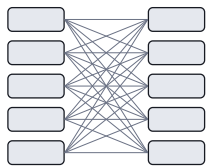
$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

# Distributed Alignment Search learns an activation subspace that maps to a concept

1. select a concept

2. rotate $a$



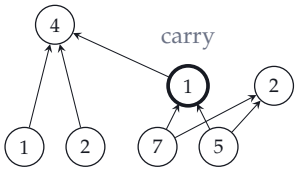




Schematic; Geiger et al. (2024); Sutter et al. (2025)

$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

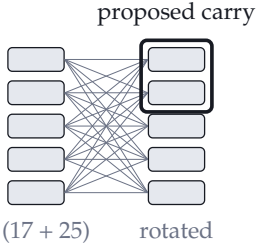
# Distributed Alignment Search learns an activation subspace that maps to a concept

1. select a concept




2. rotate $a$

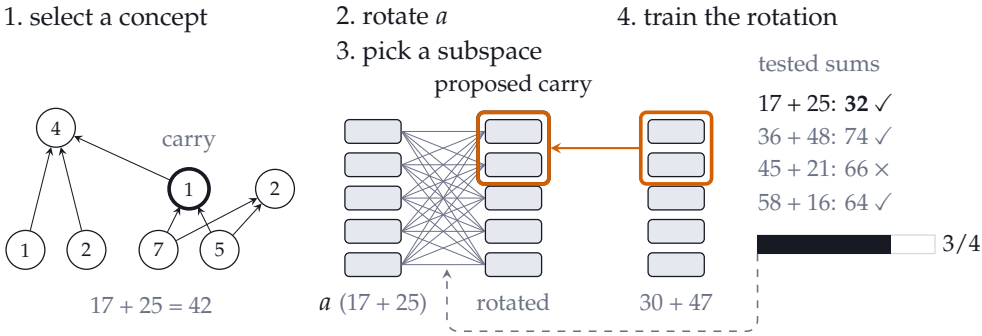
3. pick a subspace




Schematic; Geiger et al. (2024); Sutter et al. (2025)

$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

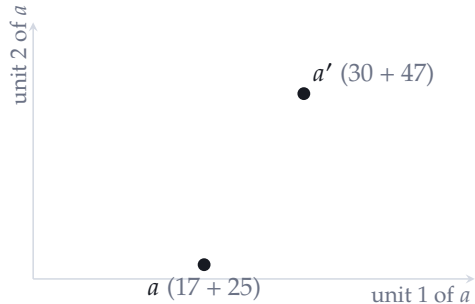
# Distributed Alignment Search learns an activation subspace that maps to a concept




Schematic; Geiger et al. (2024); Sutter et al. (2025)

$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

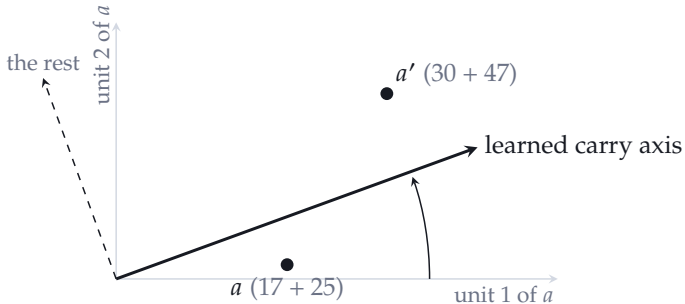
# DAS learns a rotation of the activations in which swapping a few coordinates acts like setting the carry




Schematic, two units of $a$ drawn; Geiger et al. (2024)

$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

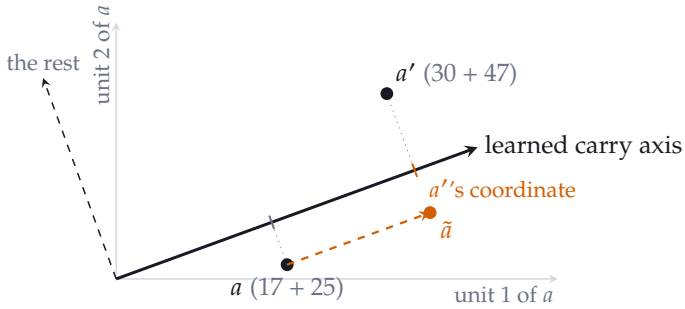
# DAS learns a rotation of the activations in which swapping a few coordinates acts like setting the carry




Schematic, two units of $a$ drawn; Geiger et al. (2024)

$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

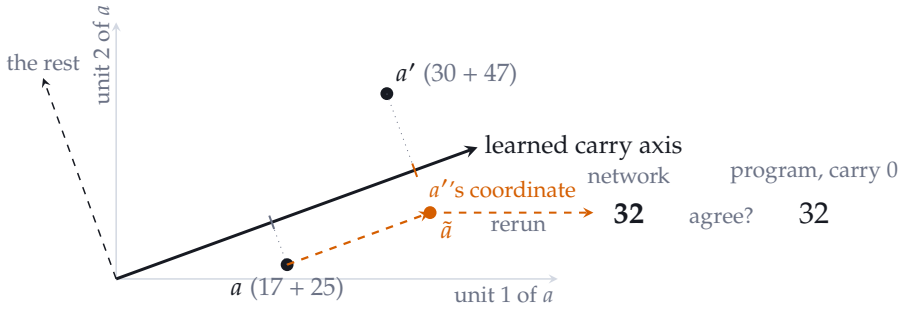
# DAS learns a rotation of the activations in which swapping a few coordinates acts like setting the carry




Schematic, two units of $a$ drawn; Geiger et al. (2024)

$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

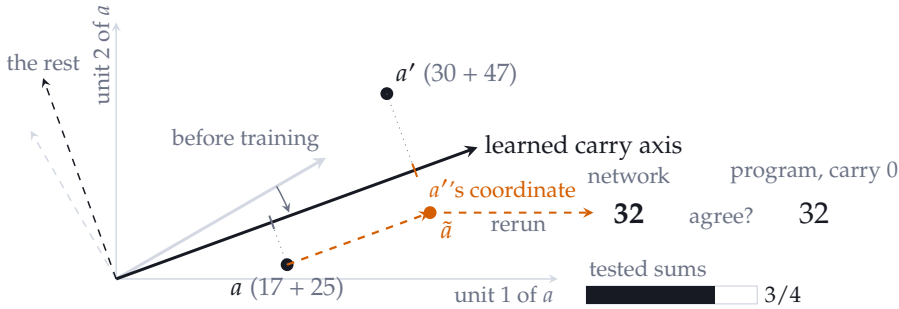
# DAS learns a rotation of the activations in which swapping a few coordinates acts like setting the carry




Schematic, two units of $a$ drawn; Geiger et al. (2024)

$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

# DAS learns a rotation of the activations in which swapping a few coordinates acts like setting the carry

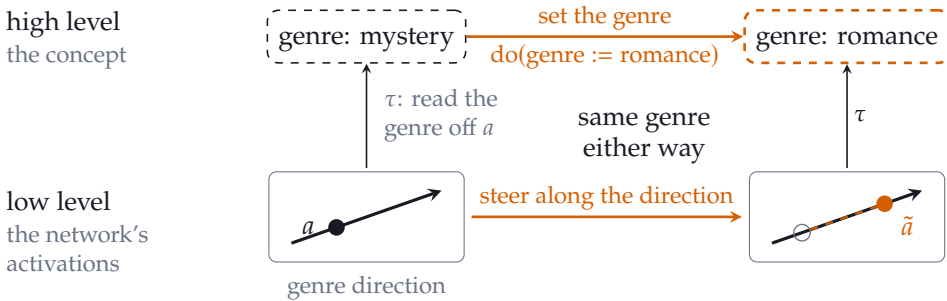



Schematic, two units of $a$ drawn; Geiger et al. (2024)

$x$ input · $a$ activation at layer $\ell$ · $y$ prediction

# A causal abstraction maps the activations to the concept

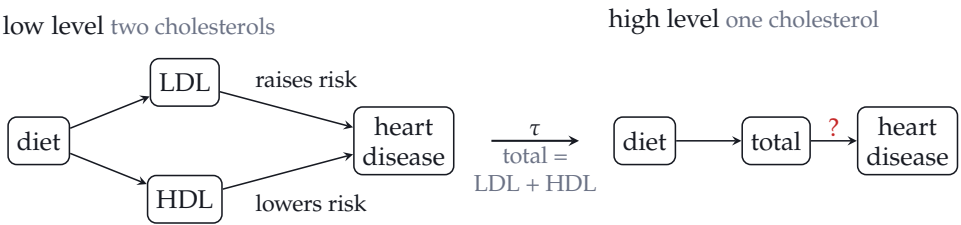
Consistency: the two routes agree.




Rubenstein et al. (2017); Beckers and Halpern (2019); Geiger et al. (2021); Geiger et al. (2025)

# Abstraction caveats

Setting the high-level variable does not say which low-level variable changes.

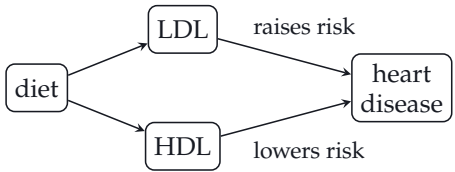



Rubenstein et al. (2017), Figure 1

# Abstraction caveats

Setting the high-level variable does not say which low-level variable changes.

low level two cholesterols




$\tau$
total = LDL + HDL

high level one cholesterol




diet A raises LDL: total up, risk up

diet B raises HDL: total up as much, risk down

in an LLM: two directions that both read as gender can move the prediction differently

Rubenstein et al. (2017), Figure 1

# Outline

Asking causal questions

causal mediation

Levels of causal questions Patrik Reizinger

the causal ladder

Identifying the answers

when are answers to causal questions unique?

# Outline

Asking causal questions
causal mediation

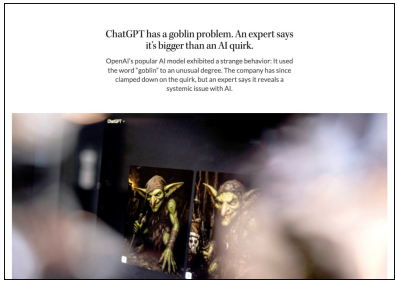
Levels of causal questions
the causal ladder

Identifying the answers Shruti Joshi
when are answers to causal questions unique?

# We keep finding unexpected LLM behaviour post hoc

Northeastern Global News, 6 May 2026; Hugging Face, 27 July 2026.

# We keep finding unexpected LLM behaviour post hoc

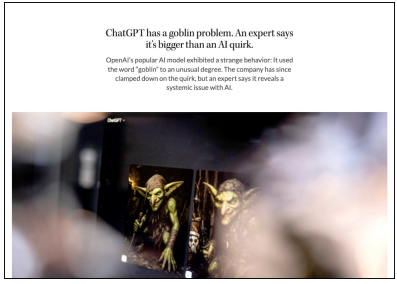


ChatGPT has a goblin problem. An expert says it's bigger than an AI quirk.

OpenAI's popular AI model exhibited a strange behavior: It used the word "goblin" to an unusual degree. The company has since clamped down on the quirk, but an expert says it reveals a systemic issue with AI.

Northeastern Global News, 6 May 2026; Hugging Face, 27 July 2026.

# We keep finding unexpected LLM behaviour post hoc



ChatGPT has a goblin problem. An expert says it's bigger than an AI quirk.

OpenAI's popular AI model exhibited a strange behavior: It used the word "goblin" to an unusual degree. The company has since clamped down on the quirk, but an expert says it reveals a systemic issue with AI.



Hugging Face

Search models, datasets, users...

Models Datasets Spaces Buckets new

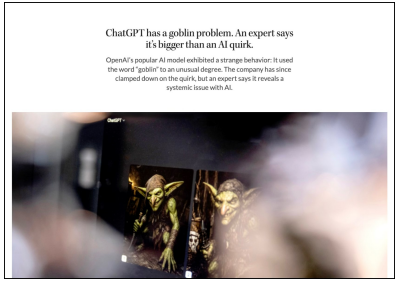
Back to Articles

**Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident**

Published July 27, 2026

Northeastern Global News, 6 May 2026; Hugging Face, 27 July 2026.

# We keep finding unexpected LLM behaviour post hoc

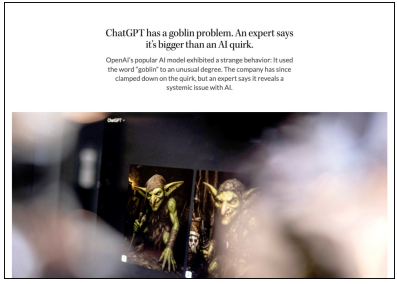



We fine-tune models,




Northeastern Global News, 6 May 2026; Hugging Face, 27 July 2026.

# We keep finding unexpected LLM behaviour post hoc




We fine-tune models,

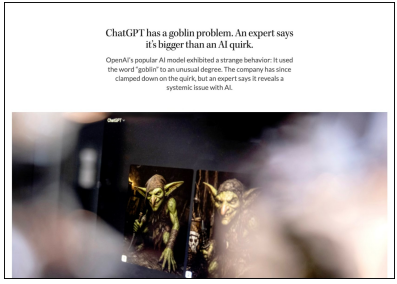
↓

find unexpected behaviours,




Northeastern Global News, 6 May 2026; Hugging Face, 27 July 2026.

# We keep finding unexpected LLM behaviour post hoc




We fine-tune models,

↓

find unexpected behaviours,

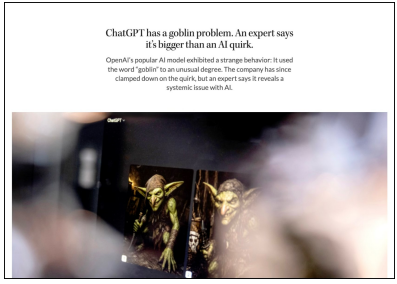
↓

fine-tune again, . . .




Northeastern Global News, 6 May 2026; Hugging Face, 27 July 2026.

# We keep finding unexpected LLM behaviour post hoc



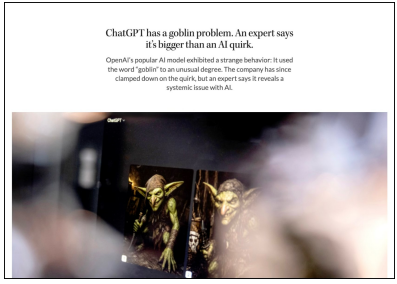

We fine-tune models,
↓
find unexpected behaviours,
↓
fine-tune again, . . .

**How do we get out of this cycle?**




Northeastern Global News, 6 May 2026; Hugging Face, 27 July 2026.

# We keep finding unexpected LLM behaviour post hoc





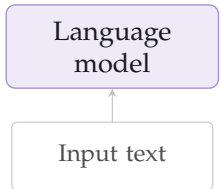


We fine-tune models,
↓
find unexpected behaviours,
↓
fine-tune again, . . .

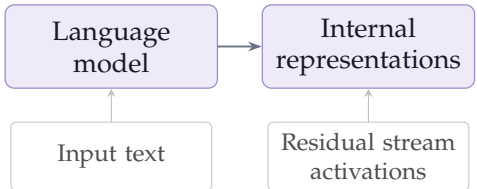
**We need ways to find changes we cannot cleanly distil into such target concepts, ideally before they show up in the outputs.**

Northeastern Global News, 6 May 2026; Hugging Face, 27 July 2026.

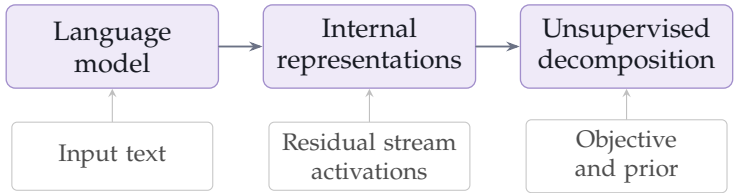
# The unsupervised interpretability pipeline



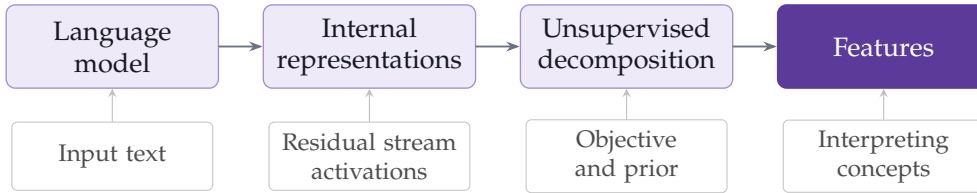
# The unsupervised interpretability pipeline



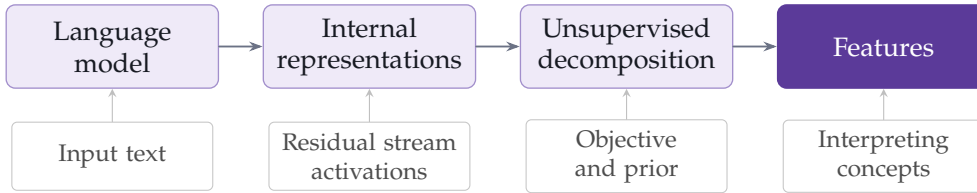
# The unsupervised interpretability pipeline



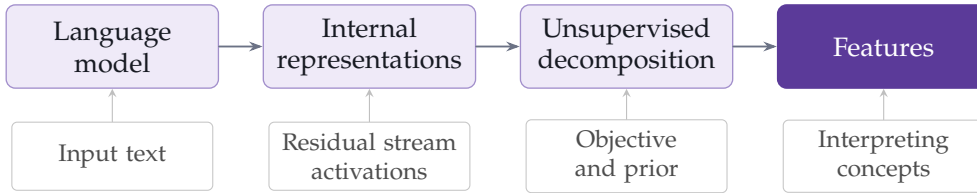
# The unsupervised interpretability pipeline



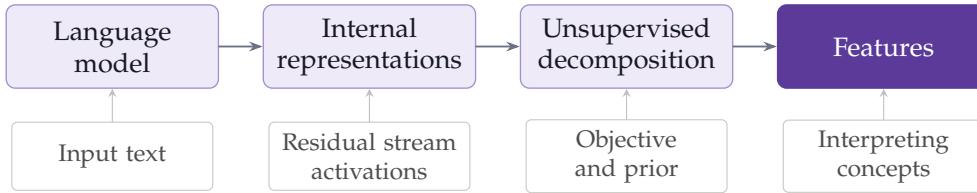
# Unsupervised interpretability requires many choices



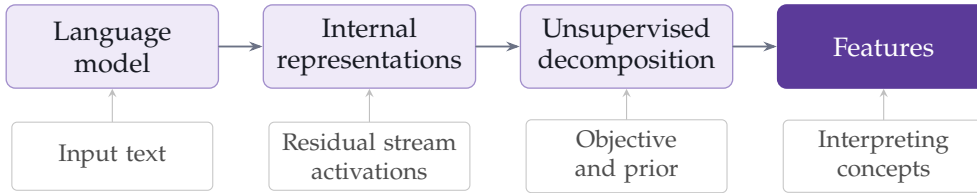
# Unsupervised interpretability requires many choices



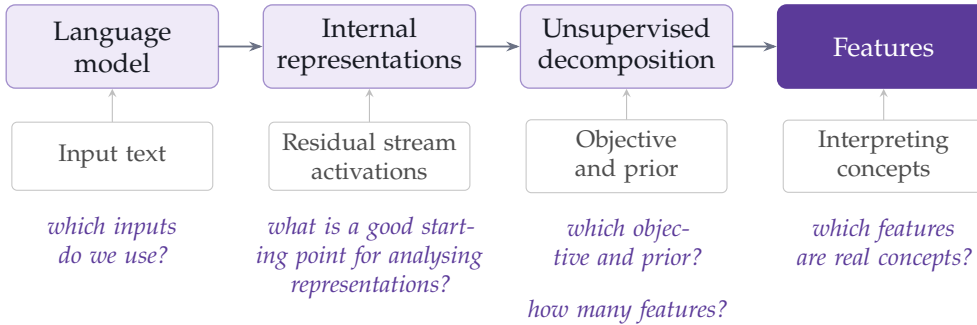
# Unsupervised interpretability requires many choices



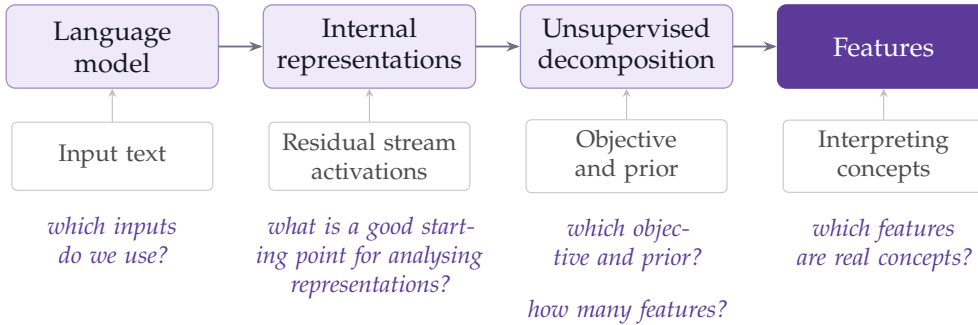
# Unsupervised interpretability requires many choices



# Unsupervised interpretability requires many choices

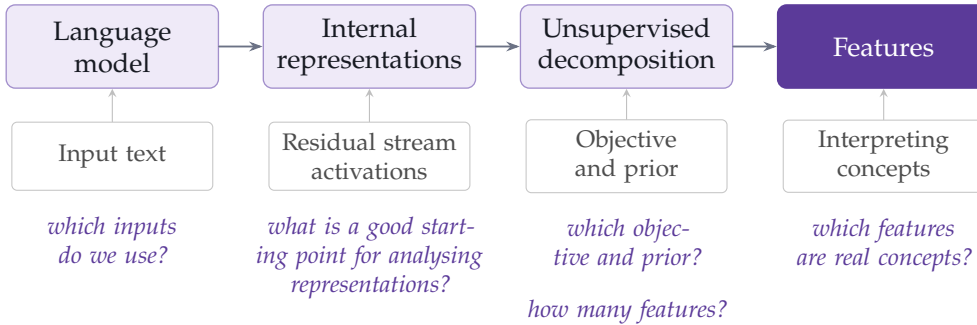


# Unsupervised interpretability requires many choices



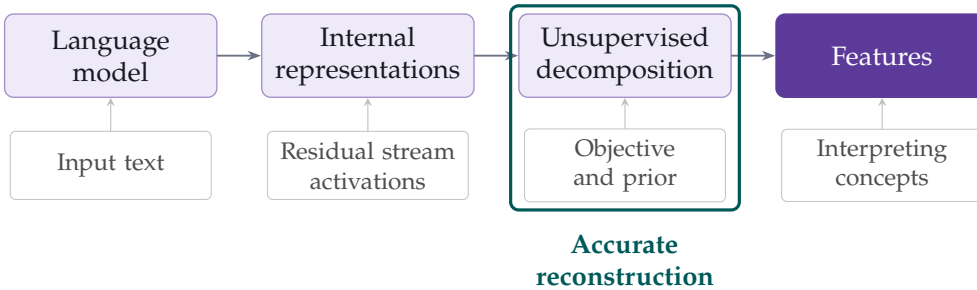

Each choice decided what concepts we can hope to recover without explicit supervision.

# Unsupervised interpretability requires many choices



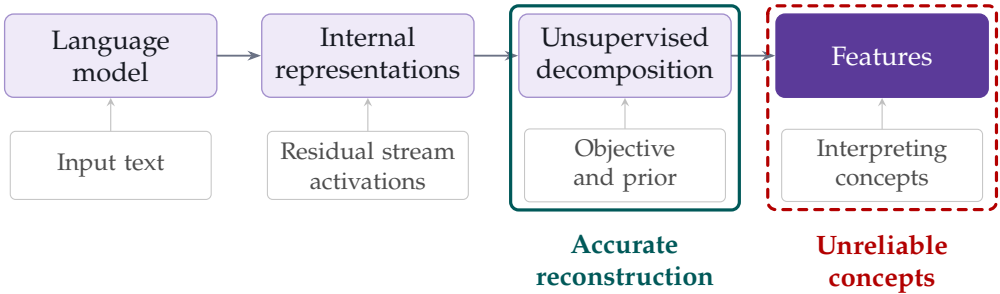

**Thus, we need reliable unsupervised interpretability such that we find consistent answers.**

# We can reconstruct activations




Sparse autoencoders reconstruct activations accurately, and their reconstruction error drops down predictably as they are scaled up. [Gao et al., 2024.]

# We can reconstruct activations, but not recover concepts




Good reconstruction does not automatically imply we have learned the concepts the model uses as several decompositions reconstruct the activations equally well. [Locatello et al., 2019. Wu et al., 2025.]

# Goal

Develop unsupervised interpretability methods that provably recover the concepts a language model uses.

# Overview

1

**Formalising concepts as latent variables**

Concepts as latent variables, and how do they produce the activations we observe?

# Overview

1

**Formalising concepts as latent variables**

Concepts as latent variables, and how do they produce the activations we observe?

2

**Learning latent variable models**

How do unsupervised methods such as PCA and sparse autoencoders learn these latent variables?

# Overview

1

**Formalising concepts as latent variables**

Concepts as latent variables, and how do they produce the activations we observe?

2

**Learning latent variable models**

How do unsupervised methods such as PCA and sparse autoencoders learn these latent variables?

3

**Learning unique solutions**

When do the learned latent variables match the model's encoded concepts?

# Overview

1

**Formalising concepts as latent variables**

Concepts as latent variables, and how do they produce the activations we observe?

2

**Learning latent variable models**

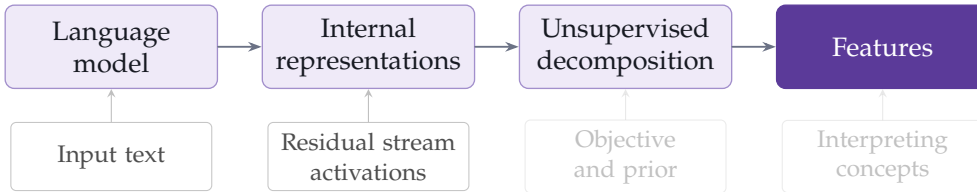
How do unsupervised methods such as PCA and sparse autoencoders learn these latent variables?

3

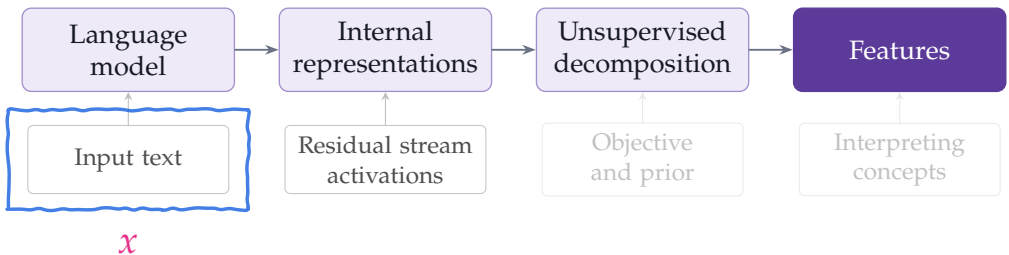
**Learning unique solutions**

When do the learned latent variables match the model's encoded concepts?

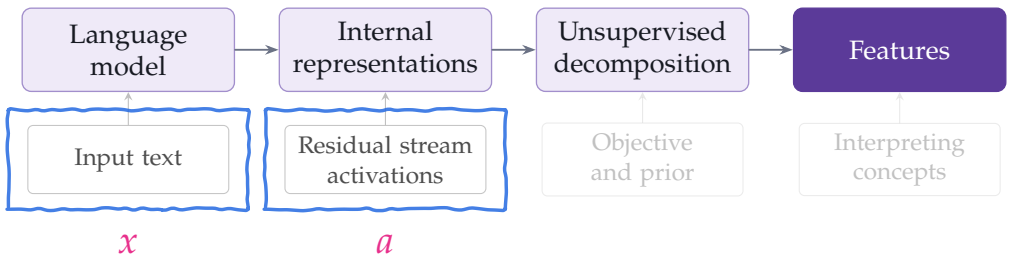
# Formalising concepts as latent variables



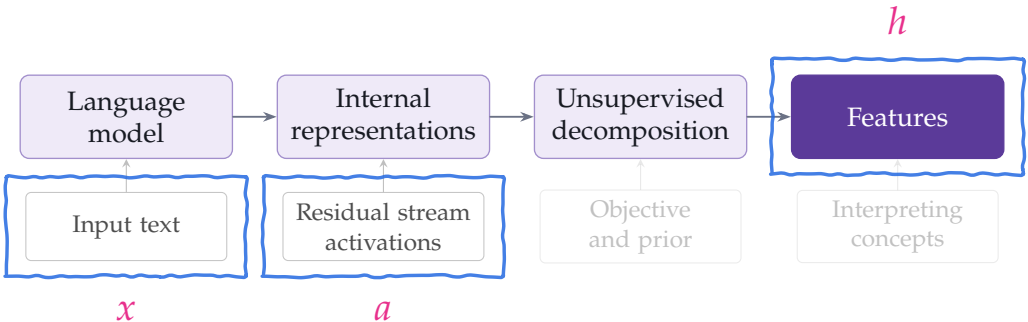
# Formalising concepts as latent variables



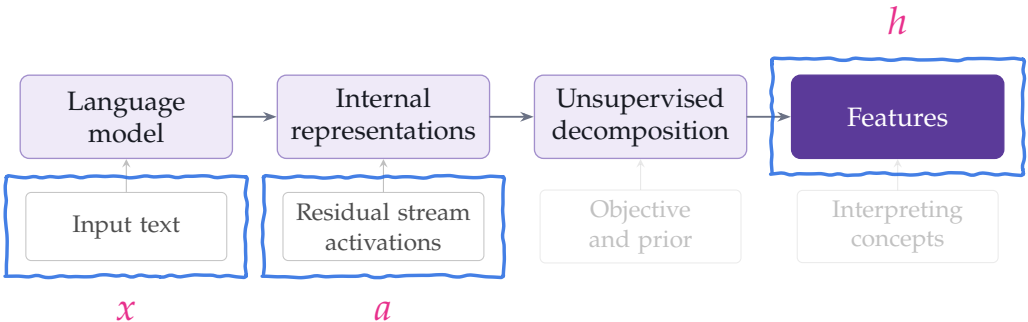
# Formalising concepts as latent variables



# Formalising concepts as latent variables

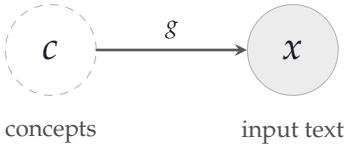


# Formalising concepts as latent variables

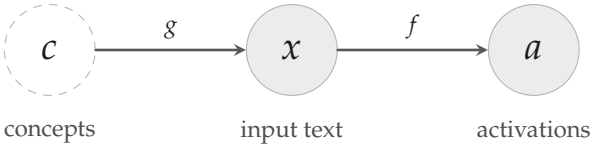



**Let $c$ denote the concepts we care about.**

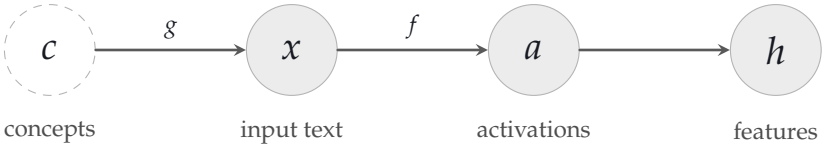
# How concepts become activations



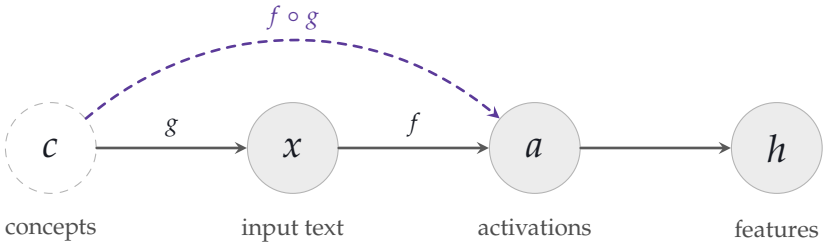
# How concepts become activations



# How concepts become activations

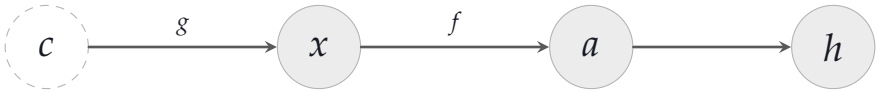


# How concepts become activations

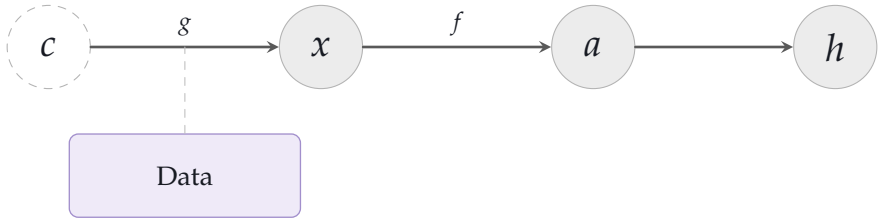



**We are interested in $f \circ g$, the map from concepts to activations.**

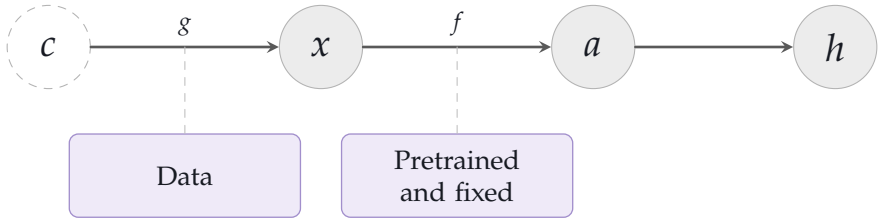
# Why model concepts this way?



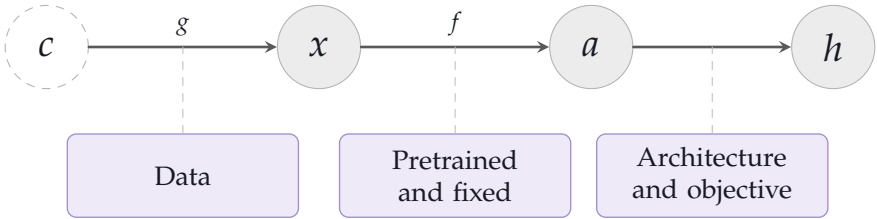
# Why model concepts this way?



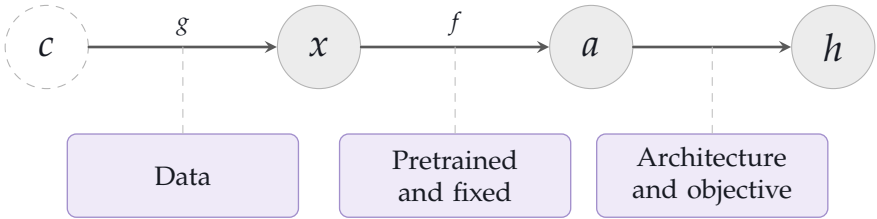
# Why model concepts this way?



# Why model concepts this way?

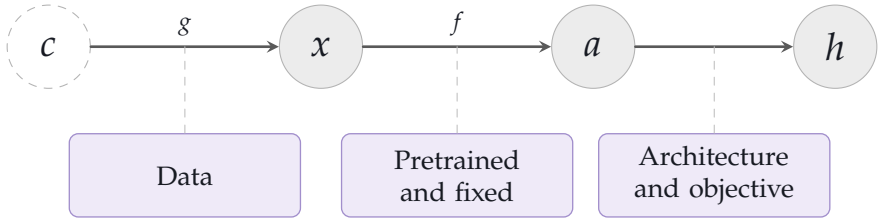


# Why model concepts this way?



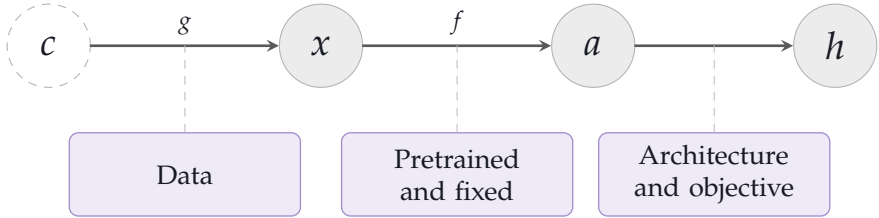

*if two concepts always occur together in the data, no method can distinguish them*

# Why model concepts this way?




Interpreting $a$ means undoing $f \circ g$, so it is tied to the data and to the architecture.

# Why model concepts this way?




**The question is, which data and architectures make concepts recoverable.**

# Overview

1

**Formalising concepts as latent variables**

Concepts as latent variables, and how do they produce the activations we observe?

2

**Learning latent variable models**

How do unsupervised methods such as PCA and sparse autoencoders learn these latent variables?

3

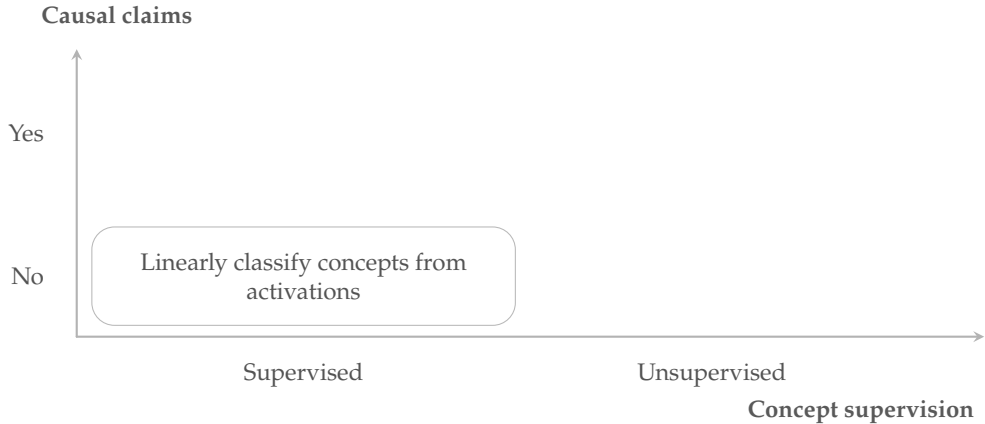
**Learning unique solutions**

When do the learned latent variables match the model's encoded concepts?

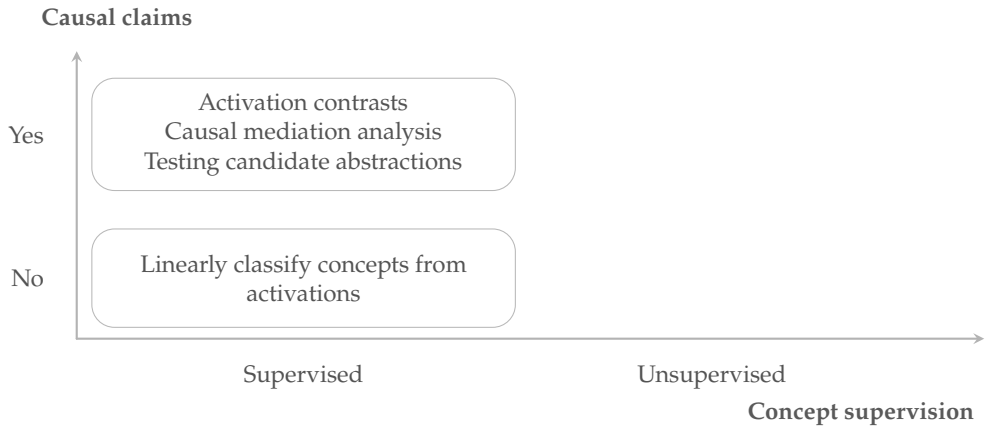
# Recall: a taxonomy of mechanistic interpretability



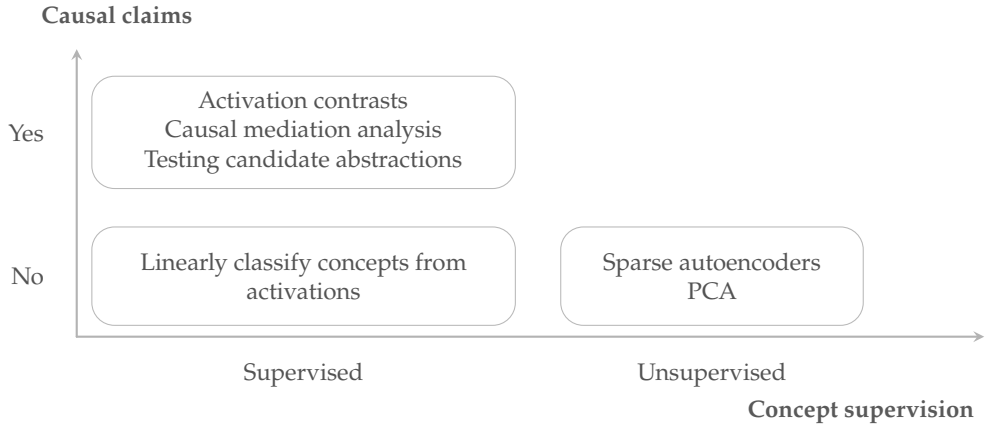
# Recall: a taxonomy of mechanistic interpretability



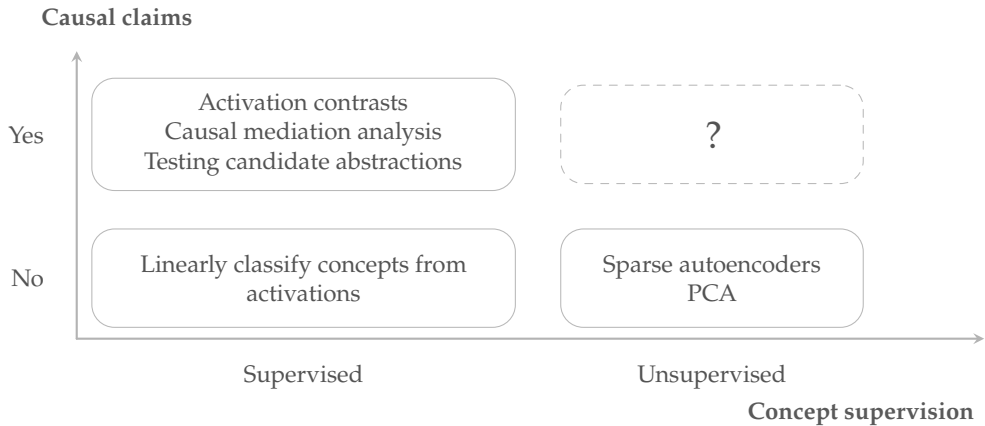
# Recall: a taxonomy of mechanistic interpretability



# Recall: a taxonomy of mechanistic interpretability

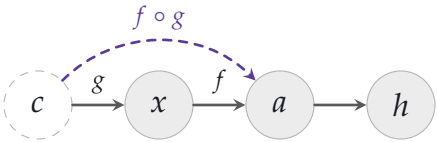


# Recall: a taxonomy of mechanistic interpretability



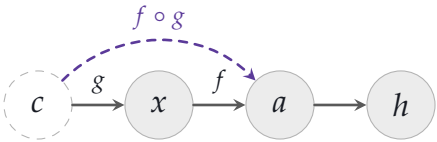
# Learning latent variable models

How the activations are generated

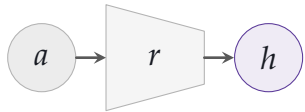


# Learning latent variable models

How the activations are generated

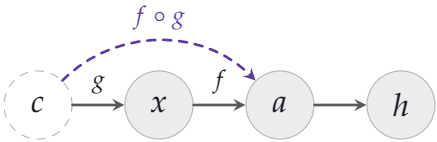



How we learn from the activations

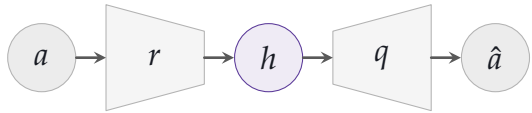


# Learning latent variable models

How the activations are generated

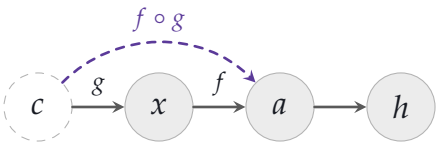



How we learn from the activations

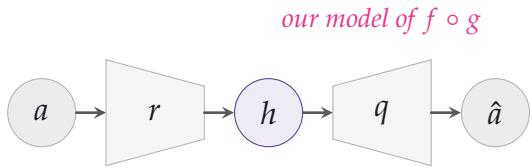


# Learning latent variable models

How the activations are generated

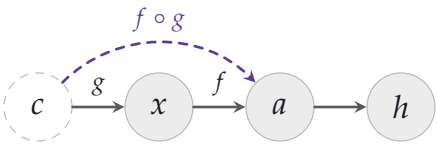



How we learn from the activations

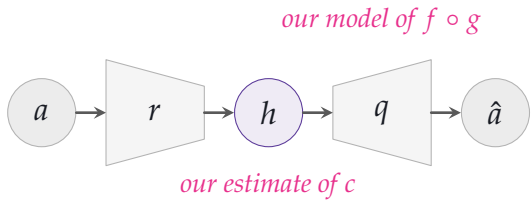


# Learning latent variable models

How the activations are generated

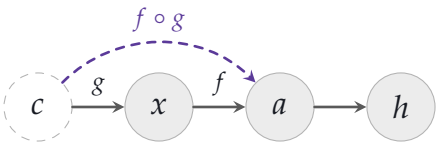



How we learn from the activations

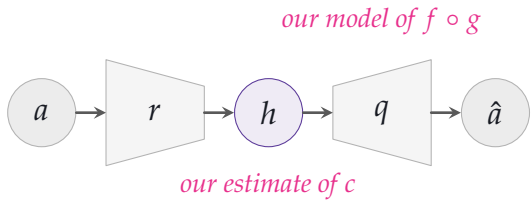


# Learning latent variable models

How the activations are generated



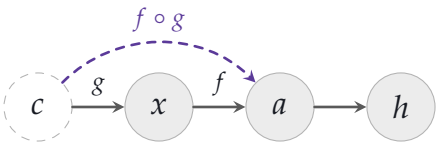

How we learn from the activations



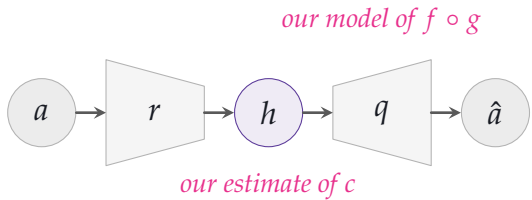

We train $r$ and $q$ so that the decoder gives back the activations, $q(r(a)) \approx a$.

# Learning latent variable models

How the activations are generated




How we learn from the activations




We train $r$ and $q$ so that the decoder gives back the activations, $q(r(a)) \approx a$.

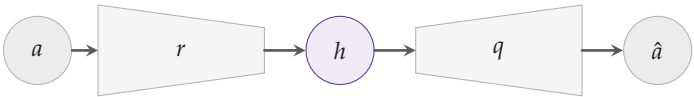
**The hope is that $h$ recovers the concepts $c$.**

# The reconstruction objective

$$\min_{r,q} \mathbb{E}\left[ \|a - q(r(a))\|^2 \right]$$

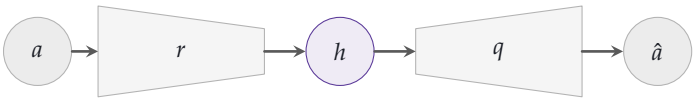
# The reconstruction objective

$$\min_{r,q} \mathbb{E}\left[ \|a - q(r(a))\|^2 \right]$$



# The reconstruction objective

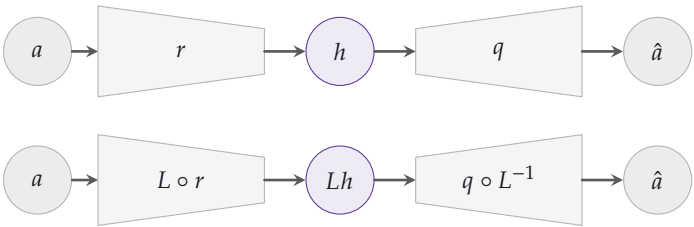
$$\min_{r,q} \mathbb{E}\left[ \|a - q(r(a))\|^2 \right]$$




*suppose h estimates c*

# The reconstruction objective

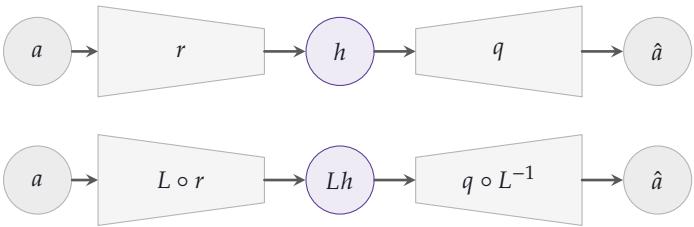
$$\min_{r,q} \mathbb{E}\left[ \|a - q(r(a))\|^2 \right]$$




*suppose $h$ estimates $c$*

# The reconstruction objective

$$\min_{r,q} \mathbb{E}\left[ \|a - q(r(a))\|^2 \right]$$

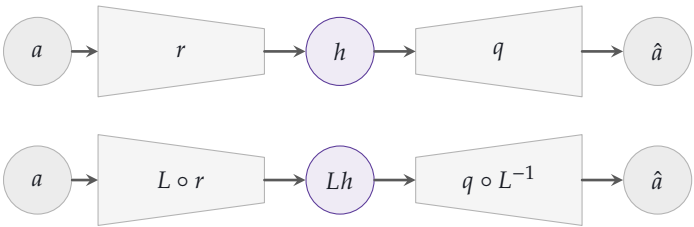



*suppose $h$ estimates $c$*

*for any invertible $L$, $q(L^{-1}Lh) = q(h)$, so the reconstruction is the same*

# The reconstruction objective

$$\min_{r,q} \mathbb{E}\left[ \|a - q(r(a))\|^2 \right]$$



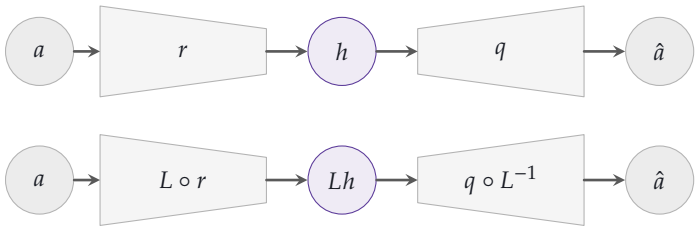

*what does Lh represent?*

*suppose h estimates c*

*for any invertible L, $q(L^{-1}Lh) = q(h)$, so the reconstruction is the same*

# The reconstruction objective

$$\min_{r,q} \mathbb{E}\left[ \|a - q(r(a))\|^2 \right]$$



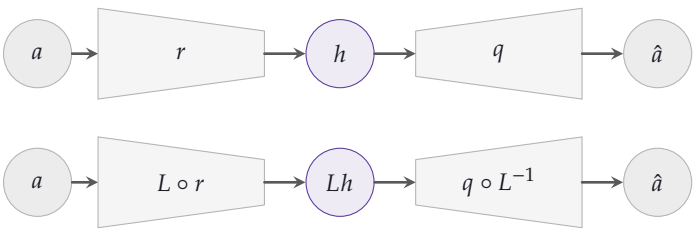

*suppose $h$ estimates $c$*

*for any invertible $L$, $q(L^{-1}Lh) = q(h)$, so the reconstruction is the same*

Each coordinate of $Lh$ mixes multiple concepts (potentially unrelated).

# The reconstruction objective

$$\min_{r,q} \mathbb{E}\left[ \|a - q(r(a))\|^2 \right]$$



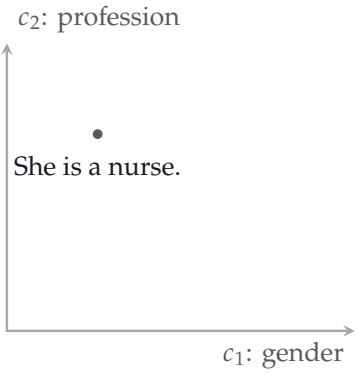

*what does $Lh$ represent?*

*suppose $h$ estimates $c$*

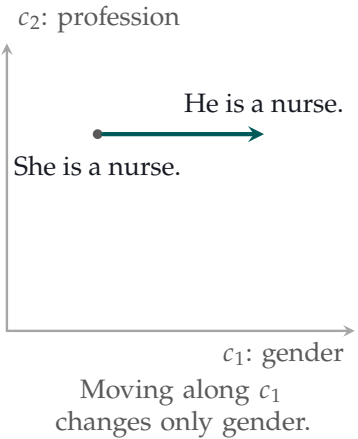
*for any invertible $L$, $q(L^{-1}Lh) = q(h)$, so the reconstruction is the same*

**Reconstruction objective alone cannot distinguish $h$ from $Lh$, such that it matches $c$.**

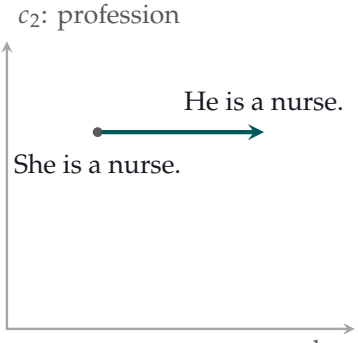
# What this means for steering



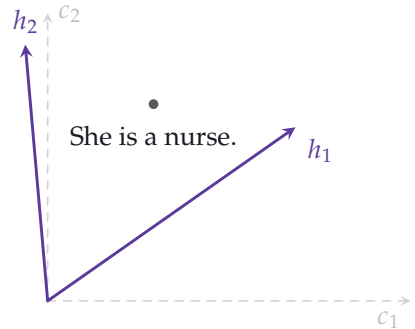
# What this means for steering



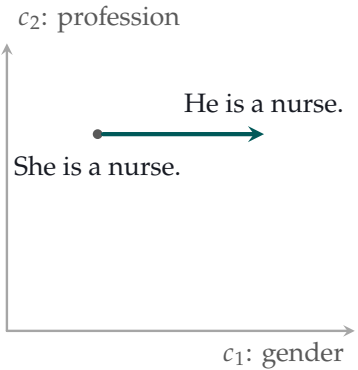
# What this means for steering



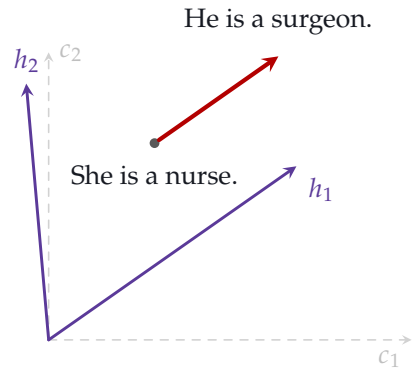

Moving along $c_1$
changes only gender.



# What this means for steering

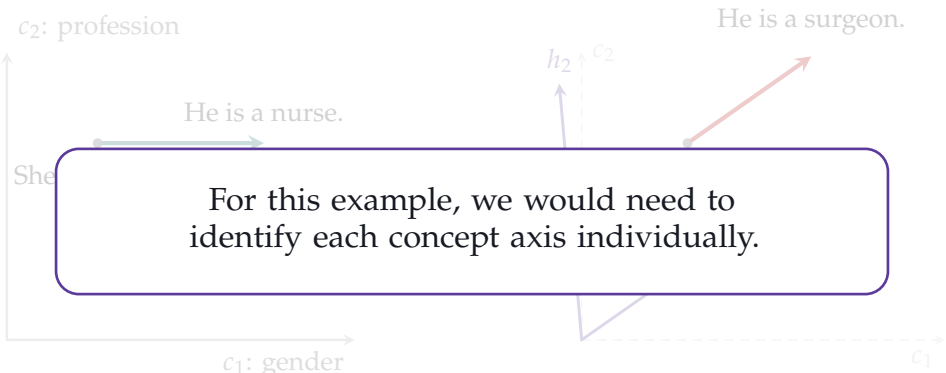



Moving along $c_1$
changes only gender.




Moving along $h_1$ changes
gender and profession at once.

# What this means for steering




For this example, we would need to identify each concept axis individually.

# Key concept: identifiable representation learning

Identifiable representation learning gives sufficient conditions, on the data and on the model, under which the $h$ we learn recovers the concepts $c$.

# Key concept: identifiable representation learning

Identifiable representation learning gives sufficient conditions, on the data and on the model, under which the $h$ we learn recovers the concepts $c$.

Reconstruction alone allows $h = L\,c$ for any invertible $L$.

# Key concept: identifiable representation learning

Identifiable representation learning gives sufficient conditions, on the data and on the model, under which the $h$ we learn recovers the concepts $c$.

Reconstruction alone allows $h = L\,c$ for any invertible $L$.

Under these conditions, $h = \underbrace{P}_{\text{permutation}} \underbrace{D}_{\text{scaling}} c$.

# Key concept: identifiable representation learning

Identifiable representation learning gives sufficient conditions, on the data and on the model, under which the $h$ we learn recovers the concepts $c$.

Reconstruction alone allows $h = L\,c$ for any invertible $L$.

Under these conditions, $h = \underbrace{P}_{\text{permutation}} \underbrace{D}_{\text{scaling}} c$.

**Neither a permutation nor a scaling mixes concepts, so each axis of $h$ aligns with a single concept.**

**So what are these conditions?**

# Overview

1

**Formalising concepts as latent variables**

Concepts as latent variables, and how do they produce the activations we observe?

2

**Learning latent variable models**

How do unsupervised methods such as PCA and sparse autoencoders learn these latent variables?

3

**Learning unique solutions**

When do the learned latent variables match the model's encoded concepts?

# Key questions

Whether we can recover the concepts, up to permutation and scaling, comes down to three questions.

# Key questions

Whether we can recover the concepts, up to permutation and scaling, comes down to three questions.

Q1 What data do we need for the answer to be unique?

# Key questions

Whether we can recover the concepts, up to permutation and scaling, comes down to three questions.

Q1 What data do we need for the answer to be unique?

Q2 What do we need to assume about the architecture and how it is trained?

# Key questions

Whether we can recover the concepts, up to permutation and scaling, comes down to three questions.

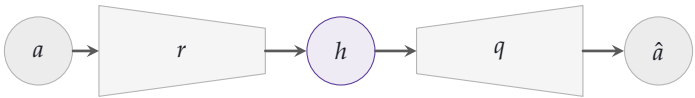
Q1 What data do we need for the answer to be unique?

Q2 What do we need to assume about the architecture and how it is trained?

Q3 How unique does the answer need to be?

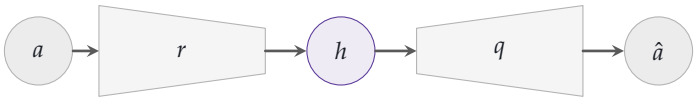
# What sparse autoencoders assume

$$\min_{r,q} \mathbb{E}\Big[ \|a - q(r(a))\|^2 \Big]$$



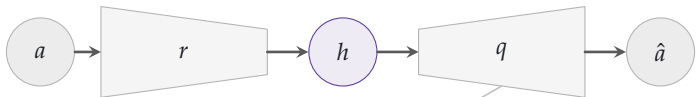
# What sparse autoencoders assume

$$\min_{r,q} \mathbb{E}\Big[ \|a - q(r(a))\|^2 + \underbrace{\lambda \|r(a)\|_1}_{\text{sparsity}} \Big]$$



# What sparse autoencoders assume

$$\min_{r,q} \mathbb{E}\Big[ \|a - q(r(a))\|^2 + \underbrace{\lambda \|r(a)\|_1}_{\text{sparsity}} \Big]$$

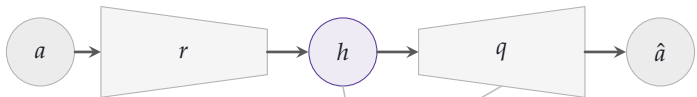



Q2

Architecture and objective: a linear decoder, and a sparsity penalty on the learned codes.

# What sparse autoencoders assume

$$\min_{r,q} \mathbb{E}\Big[ \|a - q(r(a))\|^2 + \underbrace{\lambda \|r(a)\|_1}_{\text{sparsity}} \Big]$$




Data: only a few concepts are active in each input.

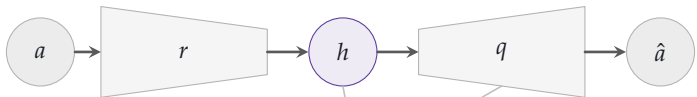
Q1

Q2

Architecture and objective: a linear decoder, and a sparsity penalty on the learned codes.

# What sparse autoencoders assume

$$\min_{r,q} \mathbb{E}\Big[ \|a - q(r(a))\|^2 + \underbrace{\lambda \|r(a)\|_1}_{\text{sparsity}} \Big]$$

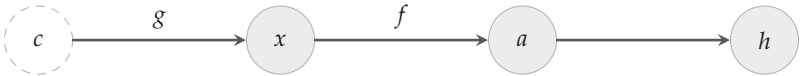



Data: only a few concepts are active in each input. **Q1**

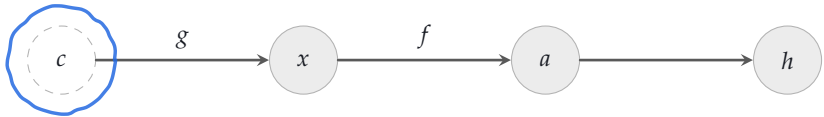
**Q2** Architecture and objective: a linear decoder, and a sparsity penalty on the learned codes.

**Sparsity rules out competing solutions only if the concepts really are sparse in the data.**

# Sparsity is an assumption about the world



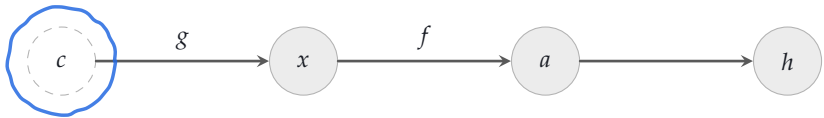
# Sparsity is an assumption about the world




*sparse autoencoders assume only a few of these are active in each input*

# Sparsity is an assumption about the world

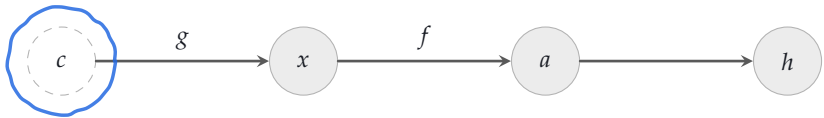
*whether that holds is a property of the world*




*sparse autoencoders assume only a few of these are active in each input*

# Sparsity is an assumption about the world

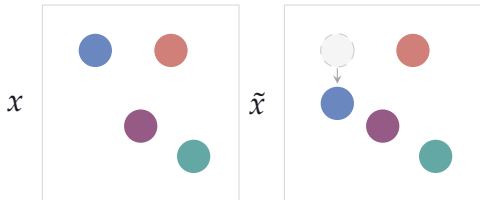
*whether that holds is a property of the world*



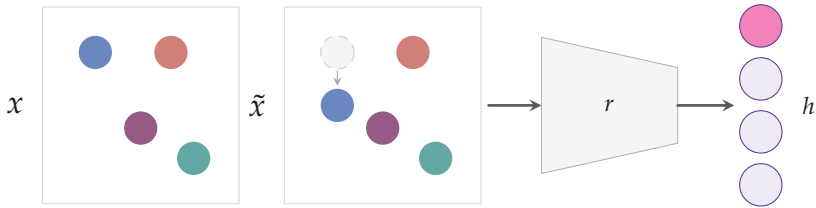

*sparse autoencoders assume only a few of these are active in each input*

**Can we move this assumption out of the world and into how we design the data?**

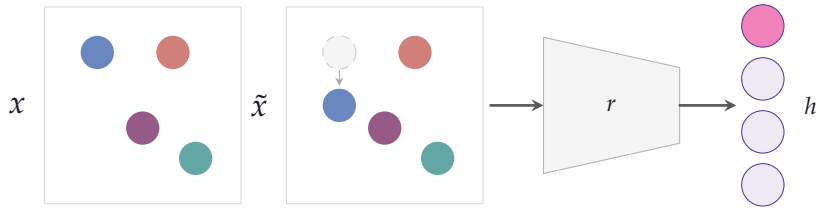
# Has identifiable representation learning studied this?



# Has identifiable representation learning studied this?



# Has identifiable representation learning studied this?



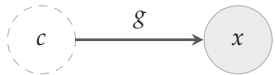

Modelling sparse or diverse interventions gives identifiable representations.

[Locatello et al., 2020. Ahuja et al., 2022. Brehmer et al., 2022. von Kügelgen et al., 2023. Zhang et al., 2023.]

# What if we look at pairs of inputs?

*(topic, tone, creature)*

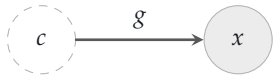
(code, playful, goblin)



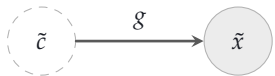

Found it! A sneaky little goblin is hiding in your loop.

# What if we look at pairs of inputs?

*(topic, tone, creature)*

(code, playful, goblin) 

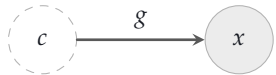

Found it! A sneaky little goblin is hiding in your loop.

(code, plain, none) 

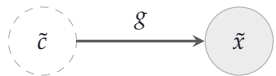

Found it. There is an off-by-one error in your loop.

# What if we look at pairs of inputs?

*(topic, tone, creature)*

(code, playful, goblin) 

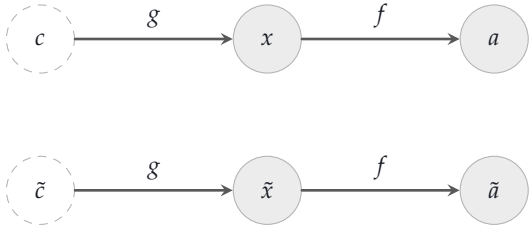

Found it! A sneaky little goblin is hiding in your loop.

(code, plain, none) 

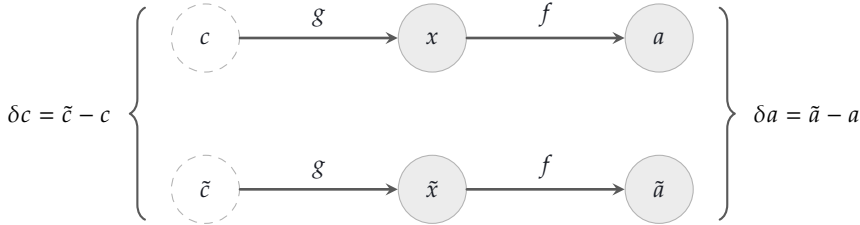

Found it. There is an off-by-one error in your loop.

Several concepts can change between the two inputs, and we never need to say which. Why look at pairs at all?

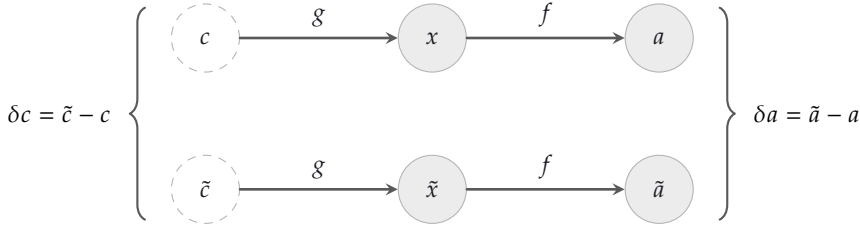
# One possible solution



# One possible solution

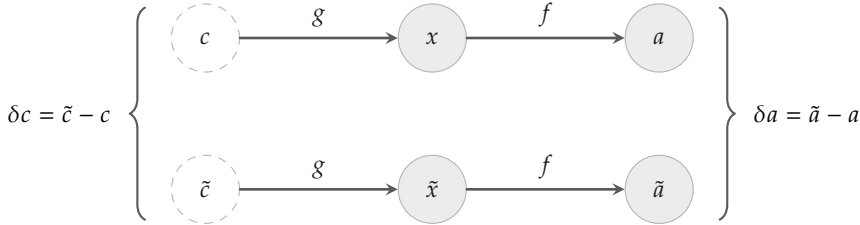


# One possible solution




Linear representation hypothesis: $a = A\,c$ for some matrix $A$.

# One possible solution




Linear representation hypothesis: $a = A\,c$ for some matrix $A$.

**Then the two differences are linearly related: $\delta a = A\,\delta c$.**

# The linear case is one of several

$\delta a$ linear in $\delta c$

*the assumption behind the linear decoder of sparse autoencoders*

# The linear case is one of several

$\delta a$ linear in $\delta c$

*the assumption behind the linear decoder of sparse autoencoders*

$\delta a$ nonlinear in $\delta c$

*also identifiable from pairs, under further conditions*

[Locatello et al., 2020. Brehmer et al., 2022. von Kügelgen et al., 2023.]

# The linear case is one of several

$\delta a$ linear in $\delta c$

*the assumption behind the linear decoder of sparse autoencoders*

$\delta a$ nonlinear in $\delta c$

*also identifiable from pairs, under further conditions*

[Locatello et al., 2020. Brehmer et al., 2022. von Kügelgen et al., 2023.]

**Here, we will assume a linear decoder, as is the standard assumption.**

# Assume the shift is sparse, not the concepts

Q1 Sparse autoencoders: only a few concepts are active in each input.
*Decided by the world.*

# Assume the shift is sparse, not the concepts

Q1 Sparse autoencoders: only a few concepts are active in each input. *Decided by the world.*

Q1 Sparse shifts: only a few concepts change within each pair, so $\delta c$ is sparse. *Decided by how we build the pairs.*

# Assume the shift is sparse, not the concepts

Q1 Sparse autoencoders: only a few concepts are active in each input. *Decided by the world.*

**Q1** Sparse shifts: only a few concepts change within each pair, so $\delta c$ is sparse. *Decided by how we build the pairs.*

We train the same autoencoder as before on $\delta a$, with the sparsity penalty on its code.

# Assume the shift is sparse, not the concepts

Q1 Sparse autoencoders: only a few concepts are active in each input. *Decided by the world.*

**Q1** Sparse shifts: only a few concepts change within each pair, so $\delta c$ is sparse. *Decided by how we build the pairs.*

We train the same autoencoder as before on $\delta a$, with the sparsity penalty on its code.

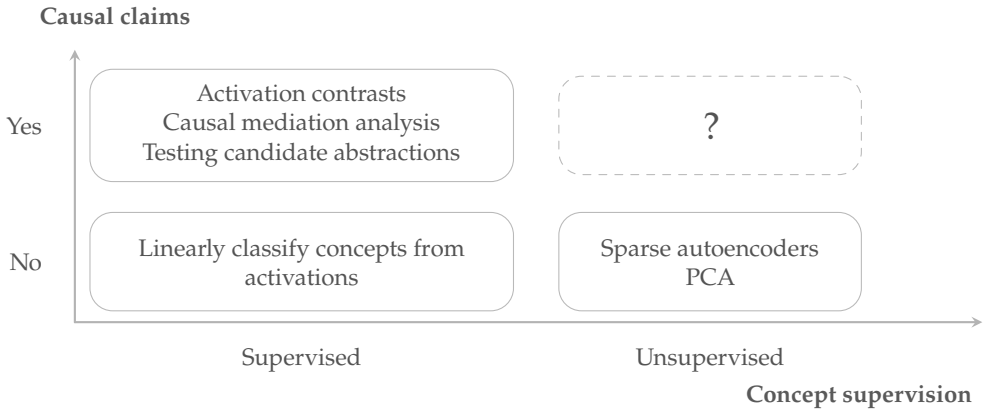
**The assumption moves from the world to experimental design.**

**Is that enough to constrain the space of solutions?**

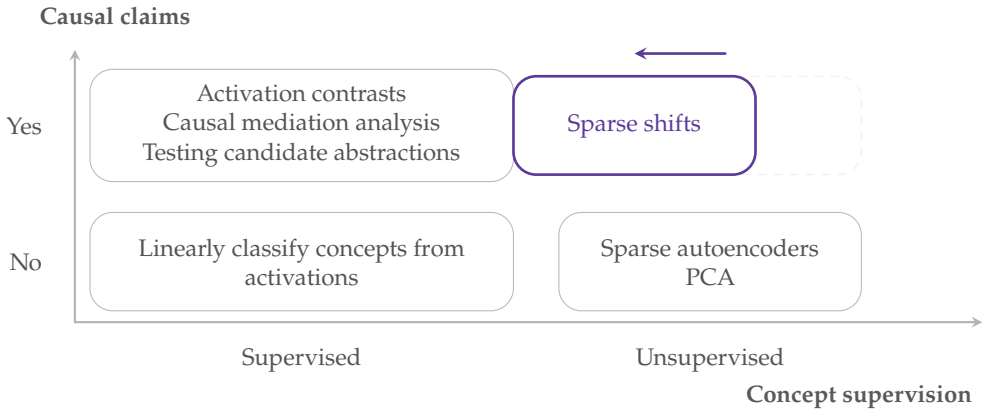
**Is that enough to constrain the space of solutions?**

Would it sit exactly where we put the question mark?

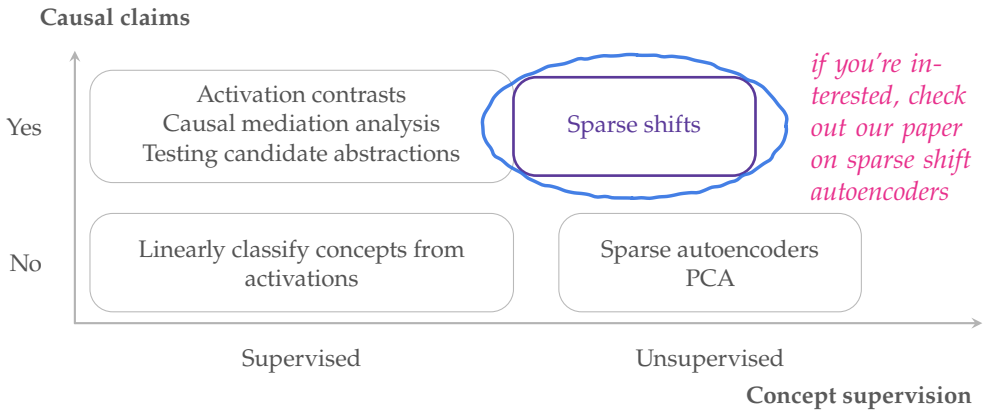
# Not quite


Causal claims
Yes
No
Activation contrasts
Causal mediation analysis
Testing candidate abstractions
?
Linearly classify concepts from activations
Sparse autoencoders
PCA
Supervised
Unsupervised
Concept supervision

# Not quite


Causal claims
Yes
No
Activation contrasts
Causal mediation analysis
Testing candidate abstractions
Sparse shifts
Linearly classify concepts from activations
Sparse autoencoders
PCA
Supervised
Unsupervised
Concept supervision

# Sparse shift autoencoders




*if you're interested, check out our paper on sparse shift autoencoders*

Joshi, Dittadi, Lachapelle and Sridhar (2025), arXiv:2502.12179.

# What flexibility does weak supervision give us?

# What flexibility does weak supervision give us?

Q1 When can we identify the concepts up to permutation and scaling?

# What flexibility does weak supervision give us?

Q1 When can we identify the concepts up to permutation and scaling?

Q2 Do we always need that much?

# Back to our key questions

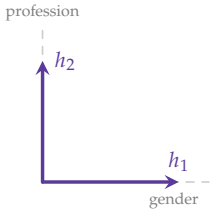
Whether we can recover the concepts, up to permutation and scaling, comes down to three questions.

Q1 What data do we need for the answer to be unique?

Q2 What do we need to assume about the architecture and how it is trained?

Q3 How unique does the answer need to be?

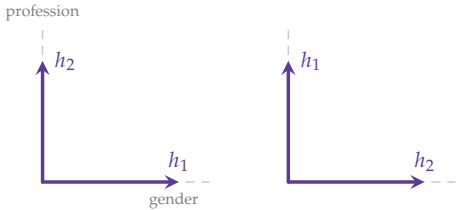
# What is an equivalence class of solutions?




the ground truth

one axis per concept

# What is an equivalence class of solutions?



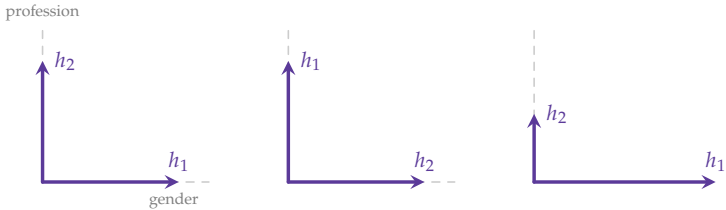

the ground truth

one axis per concept

relabelled

axes swapped

# What is an equivalence class of solutions?

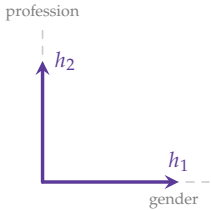



**the ground truth**
one axis per concept

**relabelled**
axes swapped

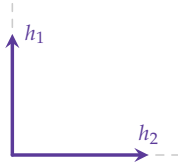
**rescaled**
different units

# What is an equivalence class of solutions?



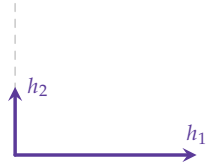

the ground truth

one axis per concept



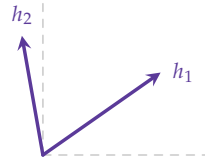

relabelled

axes swapped




rescaled

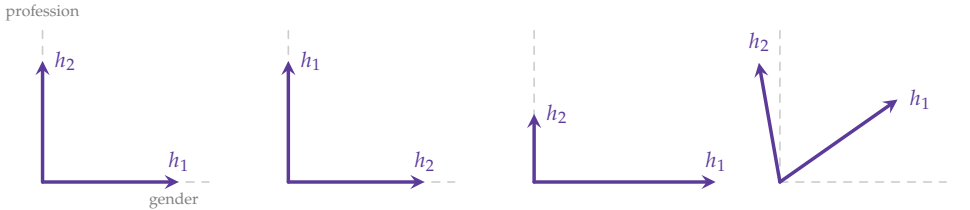
different units




linearly entangled

each axis mixes concepts

# What is an equivalence class of solutions?




**the ground truth**
one axis per concept

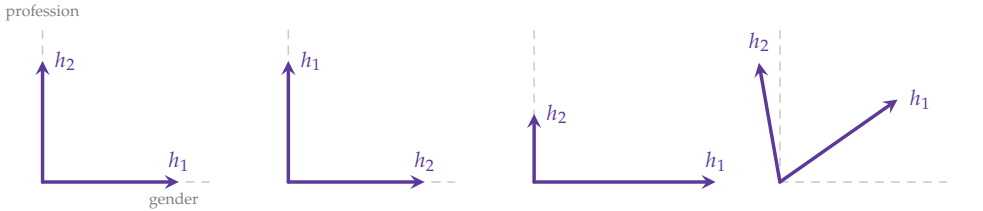
**relabelled**
axes swapped

**rescaled**
different units

**linearly entangled**
each axis mixes concepts

all four reconstruct the activations equally well

# What is an equivalence class of solutions?

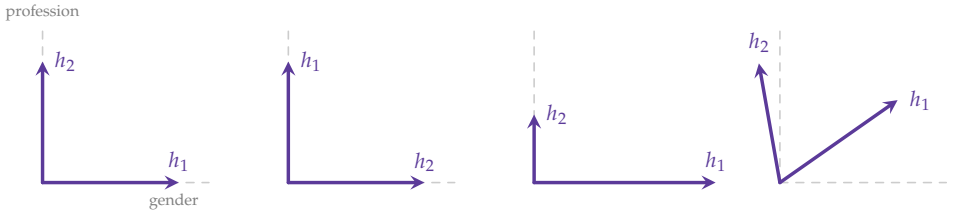



| the ground truth | relabelled | rescaled | linearly entangled |
|---|---|---|---|
| one axis per concept | axes swapped | different units | each axis mixes concepts |

all four reconstruct the activations equally well

An **equivalence class** is the set of solutions under our assumptions that our data and modeling choice cannot distinguish.

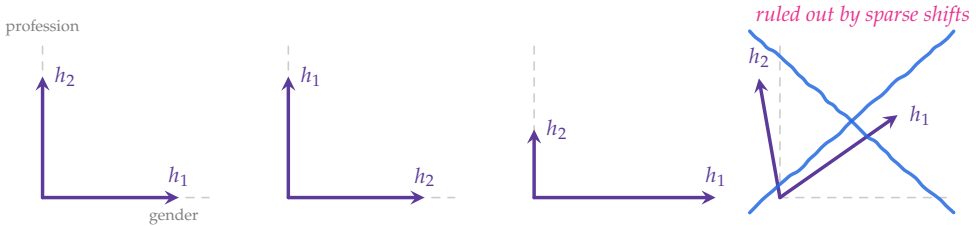
# Identifiability shrinks the equivalence class




| the ground truth | relabelled | rescaled | linearly entangled |
|---|---|---|---|
| one axis per concept | axes swapped | different units | each axis mixes concepts |

reconstruction alone: $h = Lc$ for any invertible $L$

# Identifiability shrinks the equivalence class




**the ground truth**
one axis per concept

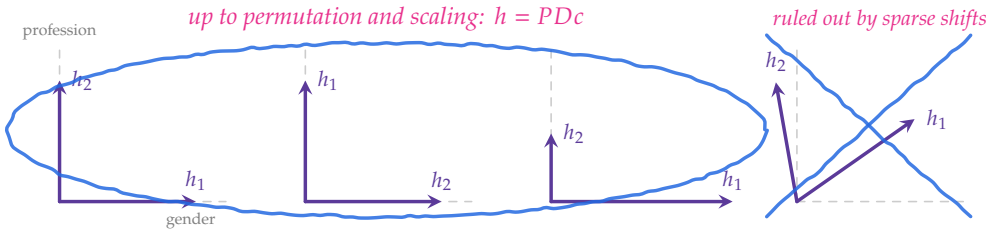
**relabelled**
axes swapped

**rescaled**
different units

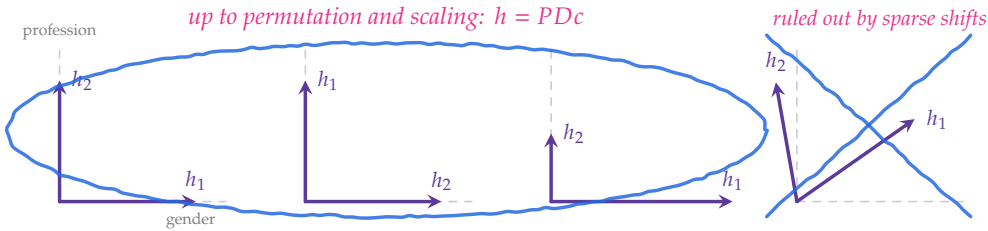
**linearly entangled**
each axis mixes concepts

reconstruction alone: $h = Lc$ for any invertible $L$

# Identifiability shrinks the equivalence class



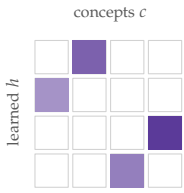
# Identifiability shrinks the equivalence class




**Every solution left has one concept per axis, so steering along any of them changes one concept.**

# How small does the class need to be?

**Permutation and scaling**

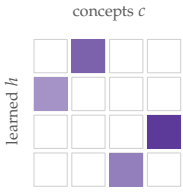



$h = PDc$

*steer or monitor a single concept*

# How small does the class need to be?

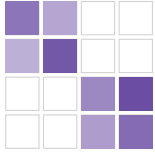
**Permutation and scaling**




$h = PDc$

*steer or monitor a single concept*

**Blocks**

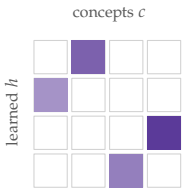


$h = Bc$, $B$ block-diagonal

*erase or isolate a group of concepts*

# How small does the class need to be?

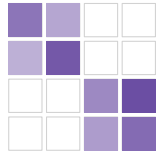
**Permutation and scaling**




$h = PDc$

*steer or monitor a single concept*

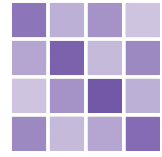
**Blocks**



$h = Bc$, $B$ block-diagonal

*erase or isolate a group of concepts*

**Any linear invertible map**



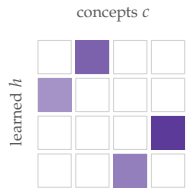
$h = Lc$

*decode concepts with a linear probe*

# How small does the class need to be?

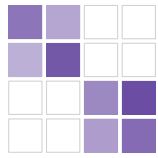
finer: needs stronger assumptions — coarser: needs weaker assumptions

**Permutation and scaling** — **Blocks** — **Any linear invertible map**



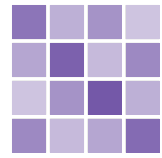

$h = PDc$

*steer or monitor a single concept*



$h = Bc$, $B$ block-diagonal

*erase or isolate a group of concepts*



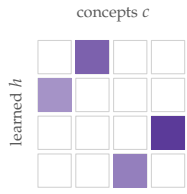
$h = Lc$

*decode concepts with a linear probe*

# How small does the class need to be?

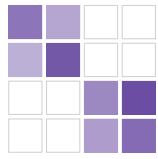
finer: needs stronger assumptions — coarser: needs weaker assumptions

**Permutation and scaling** — **Blocks** — **Any linear invertible map**



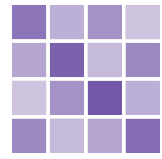

$h = PDc$

*steer or monitor a single concept*



$h = Bc$, $B$ block-diagonal

*erase or isolate a group of concepts*



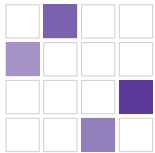
$h = Lc$

*decode concepts with a linear probe*

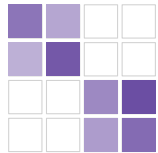
**Choose the coarsest class your problem allows: it needs the weakest assumptions.**

# Where a coarser class is enough

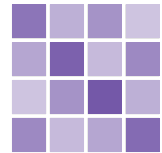
**1. Permutation and scaling**



**2. Blocks**

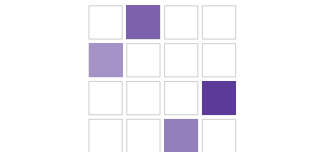


**3. Any linear invertible map**

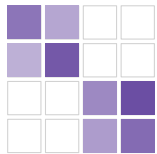


# Where a coarser class is enough

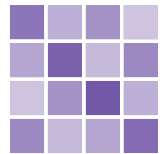
**1. Permutation and scaling**




**2. Blocks**

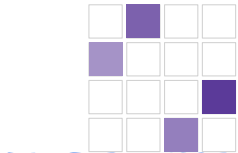


**3. Any linear invertible map**



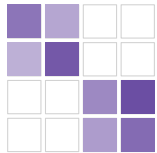
# Where a coarser class is enough

**1. Permutation and scaling**



*steer one concept at a time*

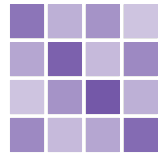
**2. Blocks**



*erase all demographic information at once*

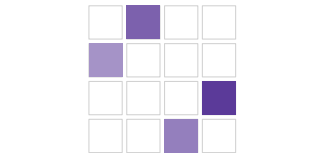
*change style while keeping content*

**3. Any linear invertible map**



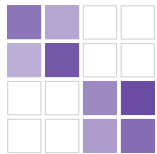
# Where a coarser class is enough

**1. Permutation and scaling**



*steer one concept at a time*

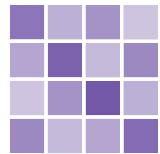
**2. Blocks**



*erase all demographic information at once*

*change style while keeping content*

**3. Any linear invertible map**



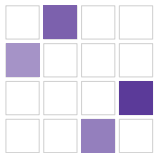
*linear probe*

*compare if two models represent the same set of concepts*

# Where a coarser class is enough

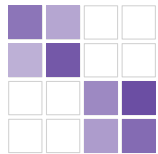
*often enough in practice*

**1. Permutation and scaling**



*steer one concept at a time*

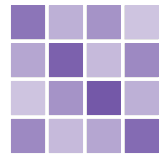
**2. Blocks**



*erase all demographic information at once*

*change style while keeping content*

**3. Any linear invertible map**



*linear probe*

*compare if two models represent the same set of concepts*

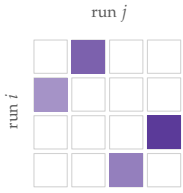
# How do different runs and models relate?

$$h_i = P_i D_i \, c, \quad h_j = P_j D_j \, c \quad \Longrightarrow \quad h_i = \underbrace{P_i D_i \, (P_j D_j)^{-1}}_{\text{a permutation and a scaling}} \quad h_j$$

# How do different runs and models relate?

$$h_i = P_i D_i \, c, \quad h_j = P_j D_j \, c \quad \Longrightarrow \quad h_i = \underbrace{P_i D_i \, (P_j D_j)^{-1}}_{\text{a permutation and a scaling}} \quad h_j$$

**Permutation and scaling**

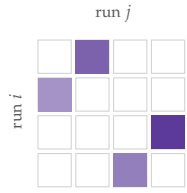



$$h_i = PD \, h_j$$

# How do different runs and models relate?

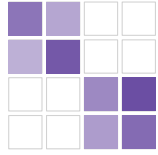
$$h_i = P_i D_i\, c, \quad h_j = P_j D_j\, c \quad \Longrightarrow \quad h_i = \underbrace{P_i D_i\, (P_j D_j)^{-1}}_{\text{a permutation and a scaling}} \quad h_j$$

**Permutation and scaling**



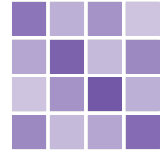

$h_i = PD\, h_j$

**Blocks**



$h_i = B\, h_j$

**Any linear invertible map**

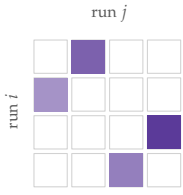


$h_i = L\, h_j$

# How do different runs and models relate?

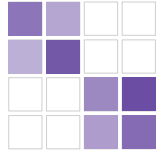
**Identifiability implies model equivalence: different runs, or different models, agree up to the class.**

**Permutation and scaling**



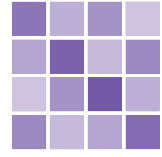

$h_i = PD\, h_j$

**Blocks**



$h_i = B\, h_j$

**Any linear invertible map**



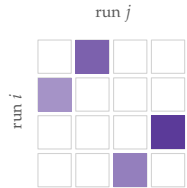
$h_i = L\, h_j$

# How do different runs and models relate?

**Identifiability implies model equivalence: different runs, or different models, agree up to the class.**

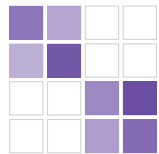
*compare models across runs with different seeds*

**Permutation and scaling**



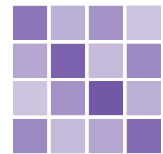

$h_i = PD\, h_j$

**Blocks**



$h_i = B\, h_j$

**Any linear invertible map**



$h_i = L\, h_j$

# Summary: why care about identifiability for interpretability?

**1** **It enables meaningful interventions.**

Intervening on a feature would change only one concept without unintended effects.

# Summary: why care about identifiability for interpretability?

**1** **It enables meaningful interventions.**

Intervening on a feature would change only one concept without unintended effects.

**2** **We can recover equivalent models.**

Different runs with different seeds would recover the same concepts, up to relabelling and rescaling.

# Summary: why care about identifiability for interpretability?

**1** **It enables meaningful interventions.**

Intervening on a feature would change only one concept without unintended effects.

**2** **We can recover equivalent models.**

Different runs with different seeds would recover the same concepts, up to relabelling and rescaling.

**3** **We can design experiments that answer causal questions uniquely!**

We can make assumptions on the data and the architecture that makes concepts recoverable.

# Summary: why care about identifiability for interpretability?

**1** **It enables meaningful interventions.**

Intervening on a feature would change only one concept without unintended effects.

**2** **We can recover equivalent models.**

Different runs with different seeds would recover the same concepts, up to relabelling and rescaling.

**3** **We can design experiments that answer causal questions uniquely!**

We can make assumptions on the data and the architecture that makes concepts recoverable.

**Identifiability is what enables us to achieve these desiderata.**

# Two takeaways

How much you can claim about the concepts an LLM uses depends on the data you learn them from and the architecture you choose.

# Two takeaways

How much you can claim about the concepts an LLM uses depends on the data you learn them from and the architecture you choose.

Where you place your assumptions matters a great deal.

# Two takeaways

How much you can claim about the concepts an LLM uses depends on the data you learn them from and the architecture you choose.

*+ how you optimise it*

Where you place your assumptions matters a great deal.

# Two takeaways

How much you can claim about the concepts an LLM uses depends on the data you learn them from and the architecture you choose.

*+ how you optimise it*

Where you place your assumptions matters a great deal.

*you can place them on the experiment rather than on the world*

# The Big Picture

## Asking causal questions

causal mediation

## Levels of causal questions

the causal ladder

## Identifying the answers

when are answers to causal questions unique?

# The Big Picture

We need experiments that enable us to distinguish between correlations and causation.

## Levels of causal questions

the causal ladder

## Identifying the answers

when are answers to causal questions unique?

# The Big Picture

We need experiments that enable us to distinguish between correlations and causation.

To understand causation, we can pose different levels of queries that give us increasingly more causal understanding of the system.

## Identifying the answers

when are answers to causal questions unique?

# The Big Picture

We need experiments that enable us to distinguish between correlations and causation.

To understand causation, we can pose different levels of queries that give us increasingly more causal understanding of the system.

We can answer these questions up to the indeterminacies determined by the assumptions encoded in our experimental design.

# Causal Foundations for Interpretability

## What Your Methods Actually Identify


| | |
|---|---|
| Shruti Joshi | Mila, Université de Montréal |
| Aaron Mueller | Boston University |
| Patrik Reizinger | Mila, Université de Montréal & ELLIS |
| Dhanya Sridhar | Mila, Université de Montréal |



EMNLP 2026 · Budapest
Wednesday 28 October 2026, 14:00–17:30 · causality4mechinterp.github.io

# *Bonus: the most important interventional questions*

# *Bonus: the most important interventional questions*

# *Opinionated Budapest recommendations*

# Eat a palacsinta, a fish soup, a fogas and a chimney cake



**Hortobágy pancake**
Hortobágyi palacsinta
Photo: Christo, CC BY-SA 4.0



**Gundel pancake**
Gundel palacsinta
Photo: Karl Gruber, CC BY-SA 3.0 AT



**Fisherman's soup**
halászlé
Photo: Lukáš Perný, CC BY-SA 4.0



**Pike-perch**
fogas (Balaton)
Photo: eLNuko, public domain



**Chimney cake**
kürtőskalács
Photo: Christo, CC BY-SA 4.0

Photos from Wikimedia Commons

# Hike Normafa, the Holdvilág gorge, Prédikálószék and the Danube Bend



**Normafa, Erzsébet tower**

Erzsébet-kilátó

Photo: Illustratedjc, CC BY-SA 4.0



**Holdvilág gorge**

Holdvilág-árok (Pilis)

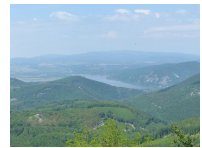
Photo: Szenti Tamás, CC BY-SA 4.0



**Prédikálószék**

Dömös

Photo: Christo, CC BY-SA 4.0



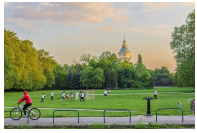
**Danube Bend**

Dunakanyar

Photo: Elekes Andor, CC BY-SA 4.0

Photos from Wikimedia Commons

# Walk Margaret Island and Gellért Hill



**Margaret Island**

Margit-sziget

Photo: Perituss, CC0



**Cave Church**

Sziklatemplom

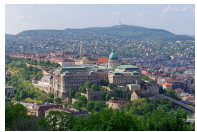
Photo: Djurika, CC BY-SA 3.0

Photos from Wikimedia Commons

# Visit Visegrád, Buda Castle, the Szentendre museum and the BME main building



**Visegrád Castle**
Fellegvár (citadel)
Photo: Szilas, public domain



**Buda Castle**
Budavári Palota, History Museum
Photo: Jakub Hałun, CC BY-SA 4.0



**Matthias Church**
Mátyás-templom
Photo: Perituss, CC0



**Szentendre open-air museum**
Skanzen, bakeries
Photo: Elekes Andor, CC BY 4.0



**BME main building**
K épület, library
Photo: Christo, CC0

Photos from Wikimedia Commons

# Pointers to the literature

## The causal ladder


Pearl and Mackenzie (2018): *The Book of Why: The New Science of Cause and Effect*

Pearl (2009): *Causality: Models, Reasoning, and Inference*

Joshi et al. (2026): *Position: Causality is Key for Interpretability Claims to Generalise*

Bareinboim et al. (2022): *On Pearl's Hierarchy and the Foundations of Causal Inference*

Feder et al. (2022): *Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond*

Huszár (2019): *Causal Inference 3: Counterfactuals*

METR (2026): *Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident*

# Pointers to the literature

## Total and direct effects

Pearl (2001): *Direct and Indirect Effects*

Wang et al. (2023): *Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small*

Pearl et al. (2016): *Causal Inference in Statistics: A Primer*

Vig et al. (2020): *Investigating Gender Bias in Language Models Using Causal Mediation Analysis*

Goldowsky-Dill et al. (2023): *Localizing Model Behavior with Path Patching*

Rushing and Nanda (2024): *Explorations of Self-Repair in Language Models*

# Pointers to the literature

## Counterfactual tokens

Oberst and Sontag (2019): *Counterfactual Off-Policy Evaluation with Gumbel-Max Structural Causal Models*

Chatzi et al. (2025): *Counterfactual Token Generation in Large Language Models*

Ravfogel et al. (2025): *Gumbel Counterfactual Generation From Language Models*

# Pointers to the literature

## Removing a direction, and counterfactuals about concepts

Ravfogel et al. (2020): *Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection*

Elazar et al. (2021): *Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals*

Kumar et al. (2022): *Probing Classifiers are Unreliable for Concept Removal and Detection*

Li et al. (2023): *Inference-Time Intervention: Eliciting Truthful Answers from a Language Model*

Rubenstein et al. (2017): *Causal Consistency of Structural Equation Models*

Beckers and Halpern (2019): *Abstracting Causal Models*

Geiger et al. (2024): *Finding Alignments Between Interpretable Causal Variables and Distributed Neural Representations*

# Pointers to the literature

## Removing a direction, and counterfactuals about concepts


Geiger et al. (2021): *Causal Abstractions of Neural Networks*

Geiger et al. (2025): *Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability*

Geiger et al. (2022): *Inducing Causal Structure for Interpretable Neural Networks*

Wu et al. (2023): *Interpretability at Scale: Identifying Causal Mechanisms in Alpaca*

Huang et al. (2024): *RAVEL: Evaluating Interpretability Methods on Disentangling Language Model Representations*

Sutter et al. (2025): *The Non-Linear Representation Dilemma: Is Causal Abstraction Enough for Mechanistic Interpretability?*


The full bibliography is on the tutorial website.