Evaluation and Interpretability, Pt. II

(Mechanistic) Interpretability

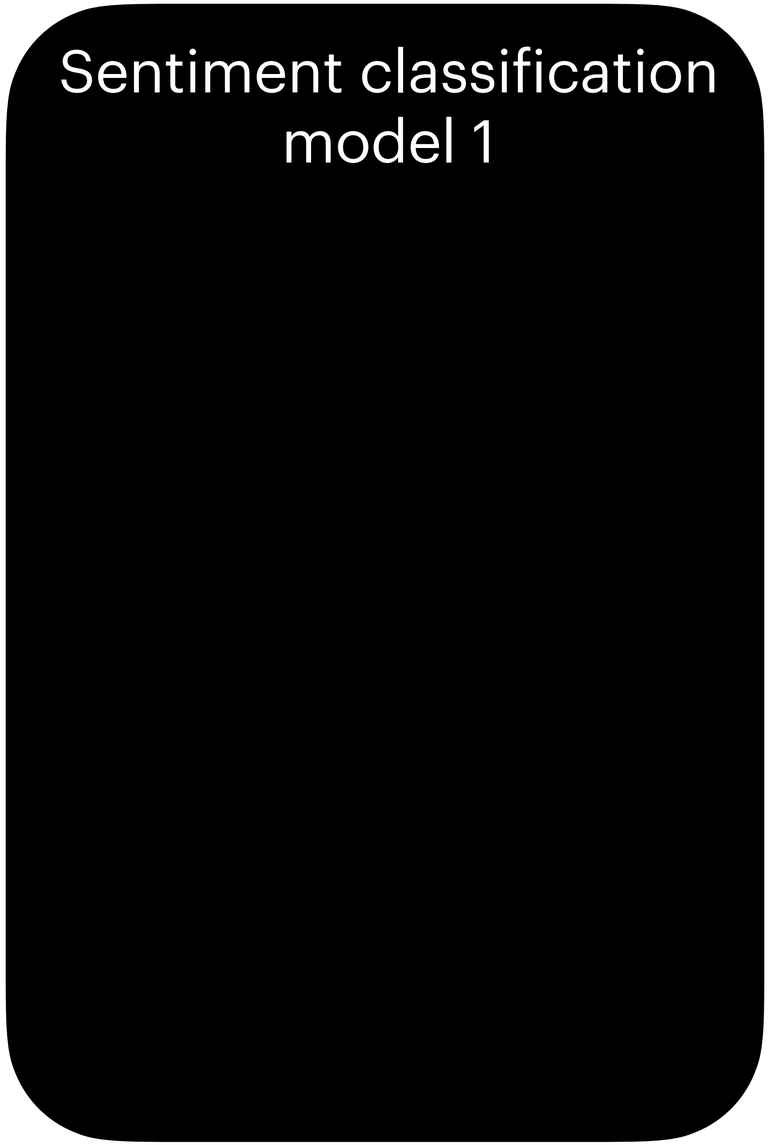
Aaron Mueller

CAS CS 505: Introduction to Natural Language Processing

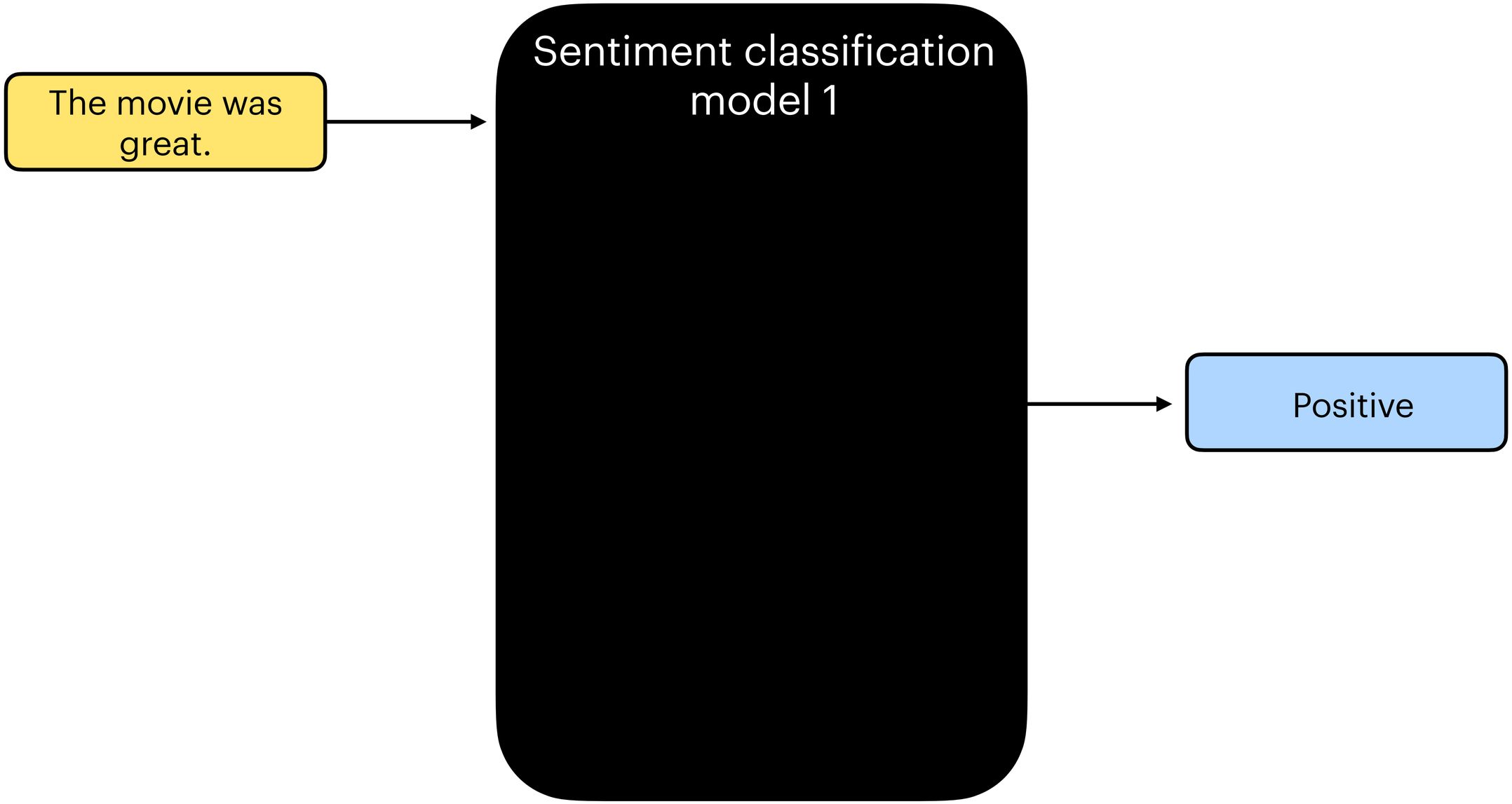
Spring 2026

Boston University

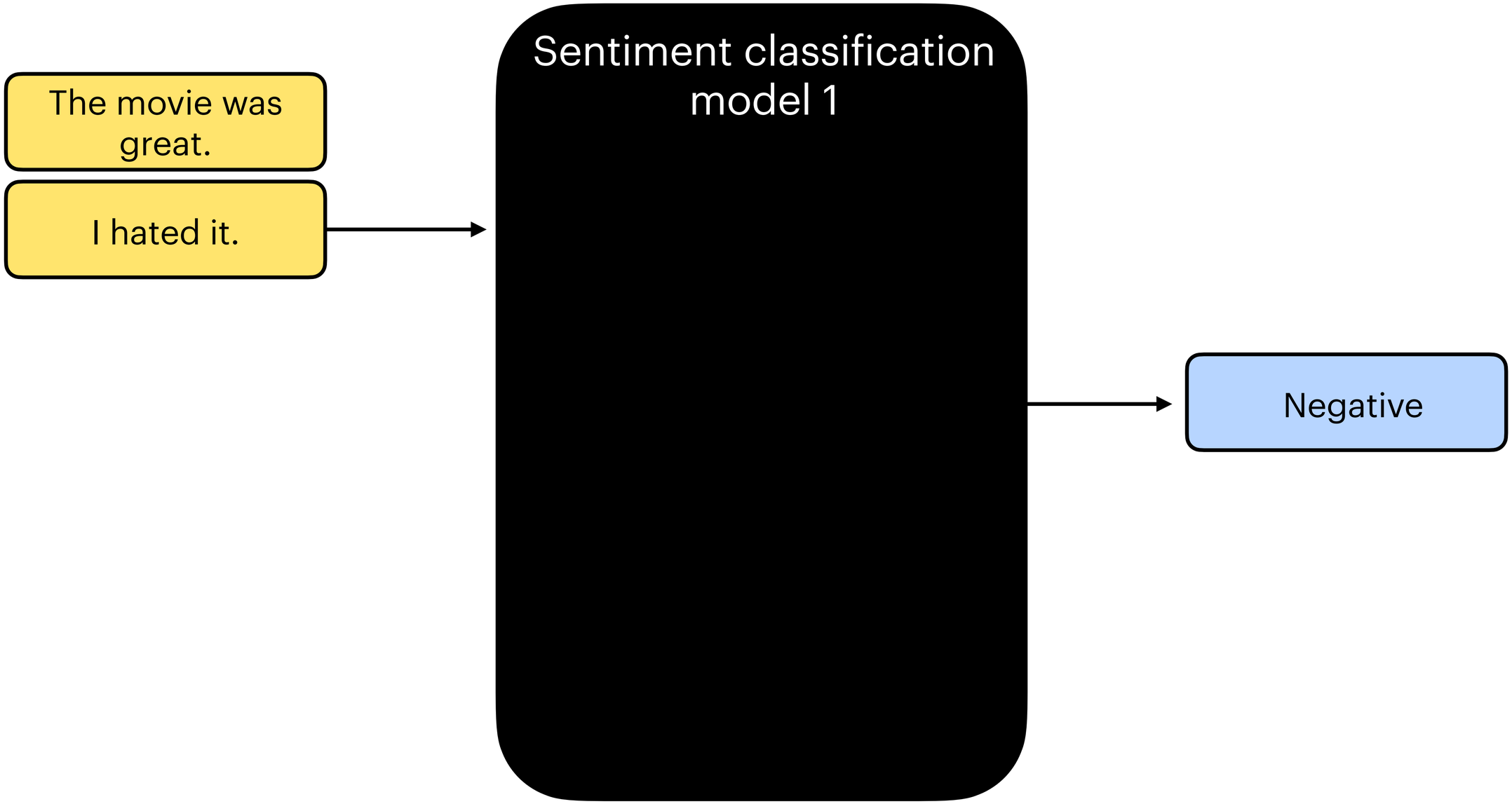
The Limitations of Behavioral Testing



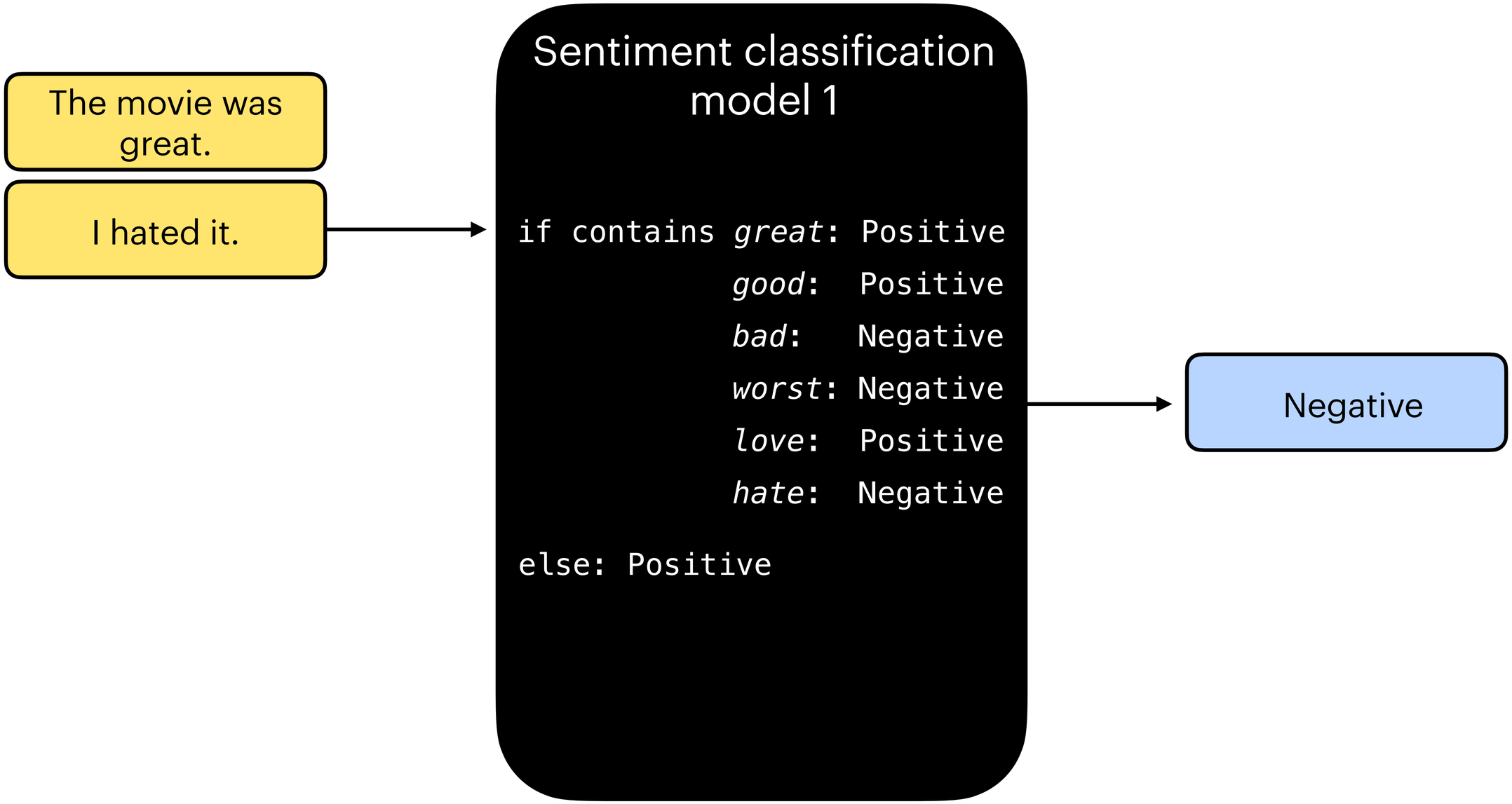
The Limitations of Behavioral Testing



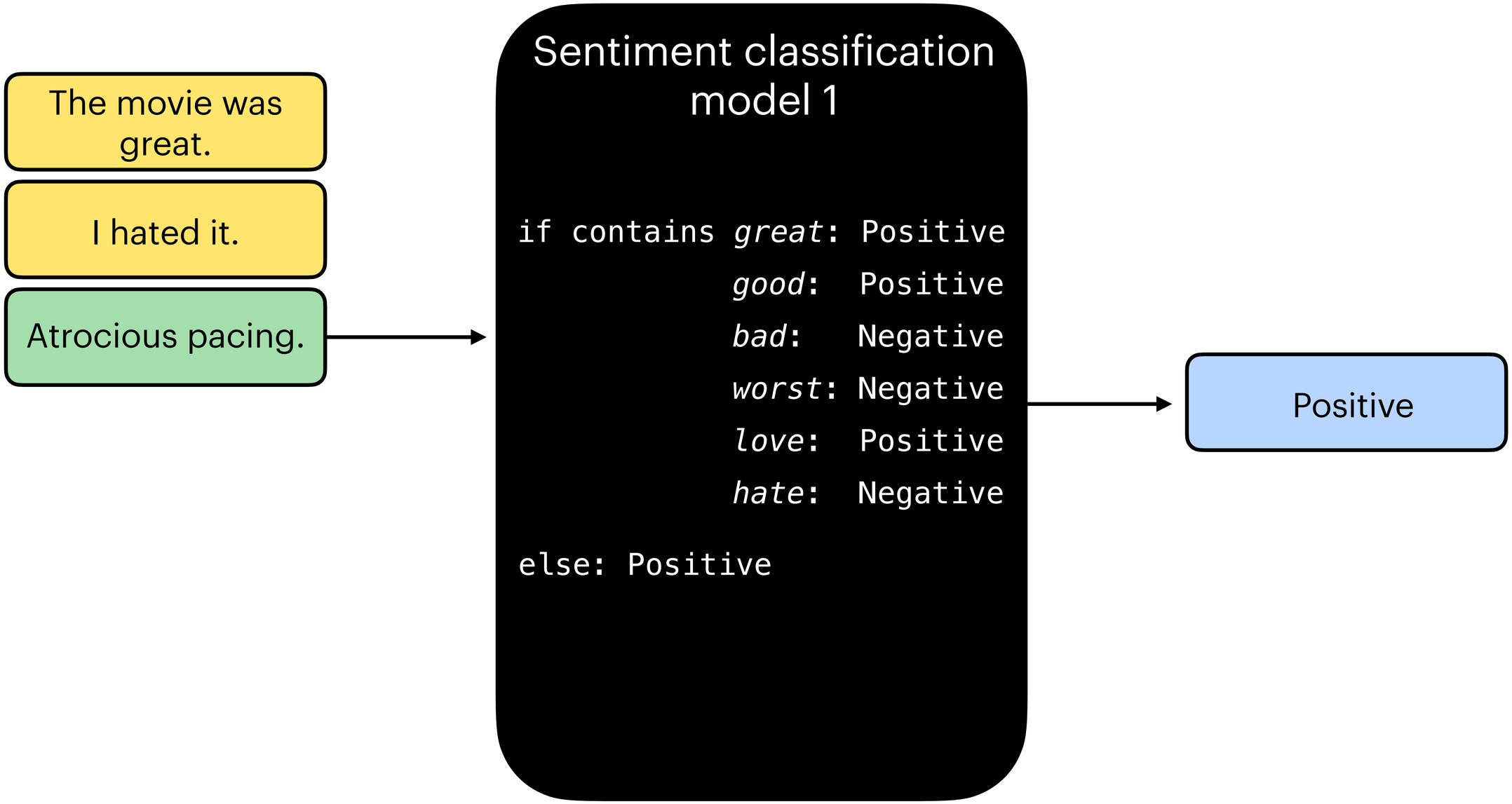
The Limitations of Behavioral Testing



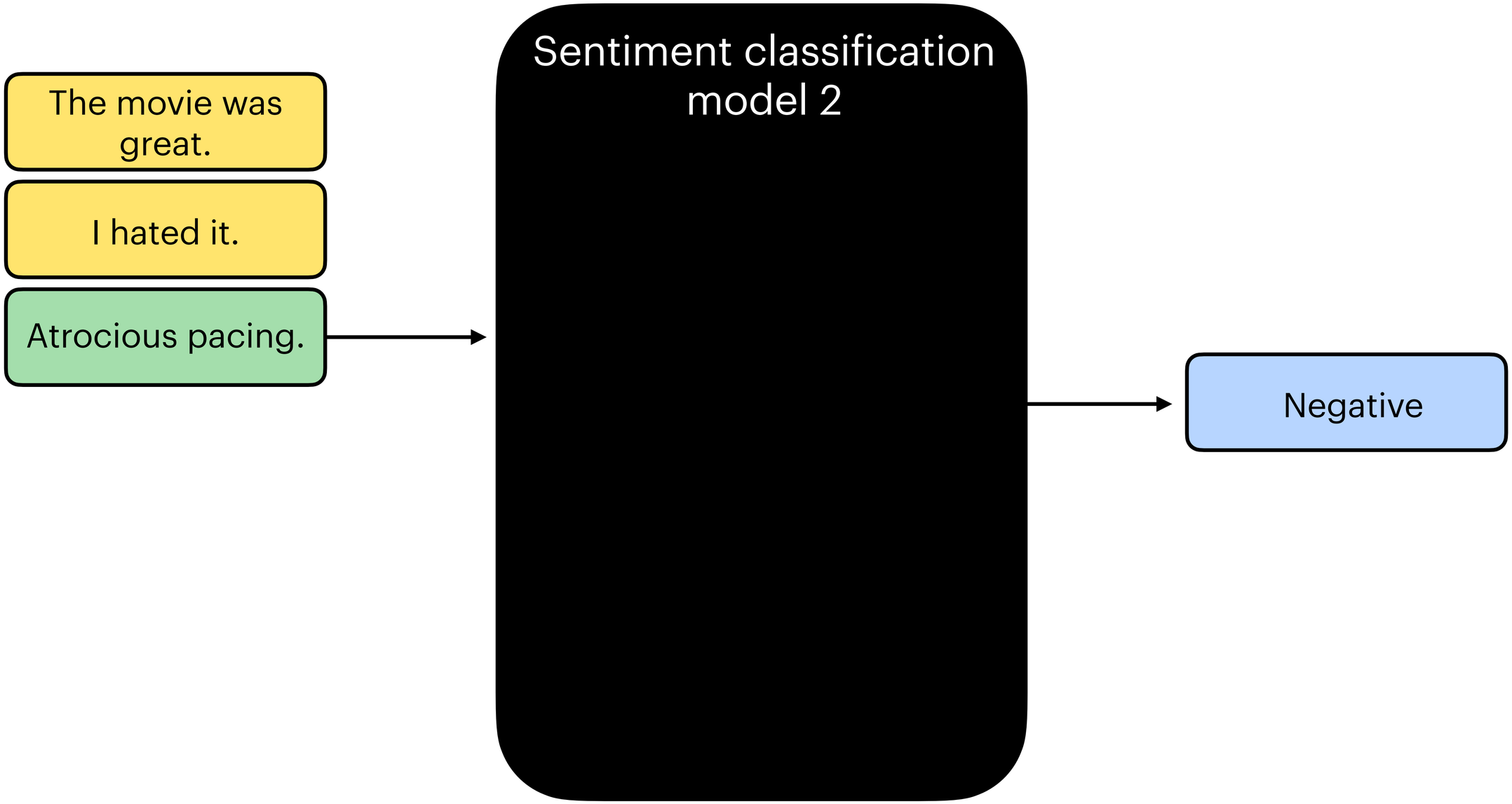
The Limitations of Behavioral Testing



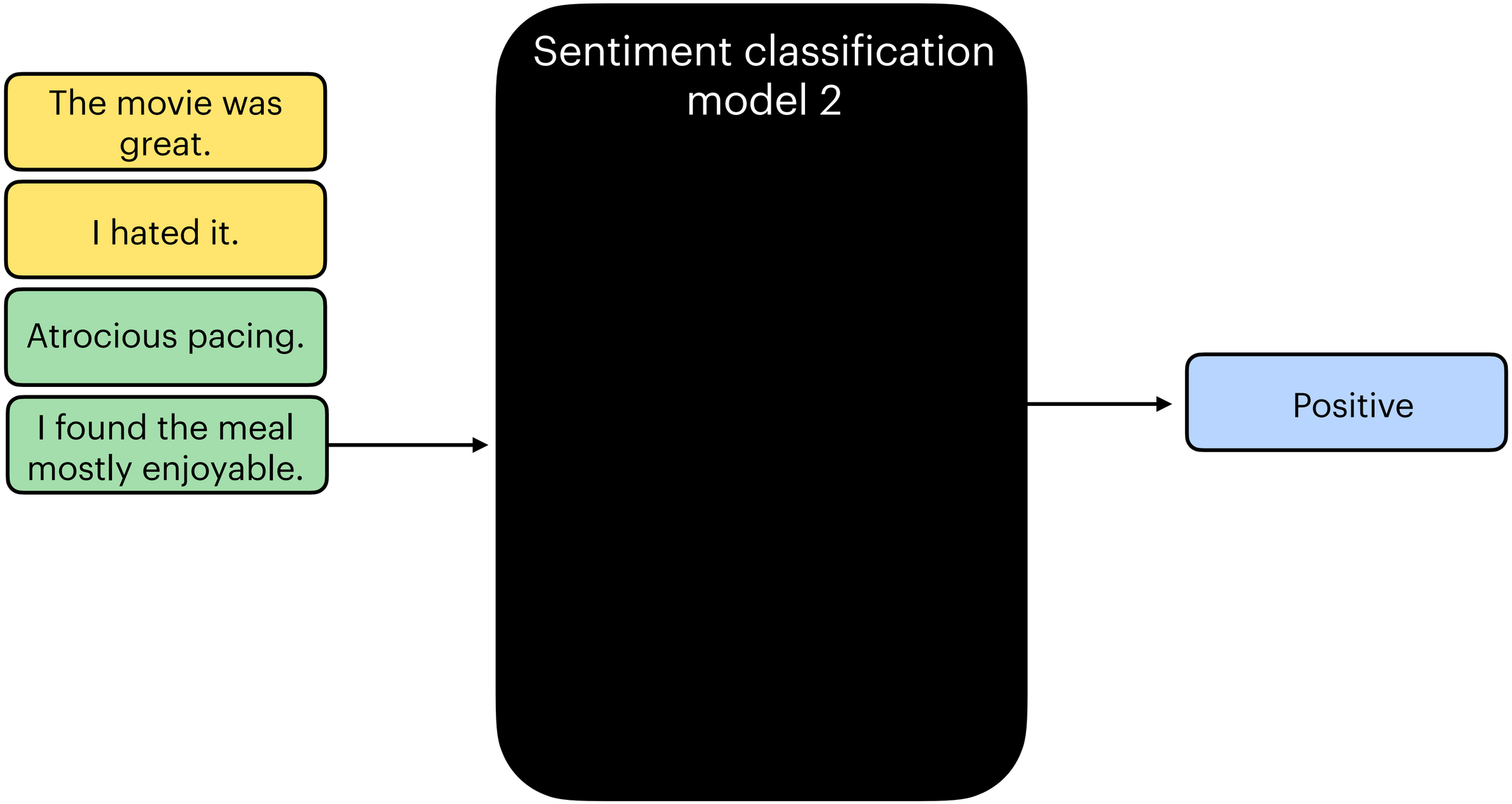
The Limitations of Behavioral Testing



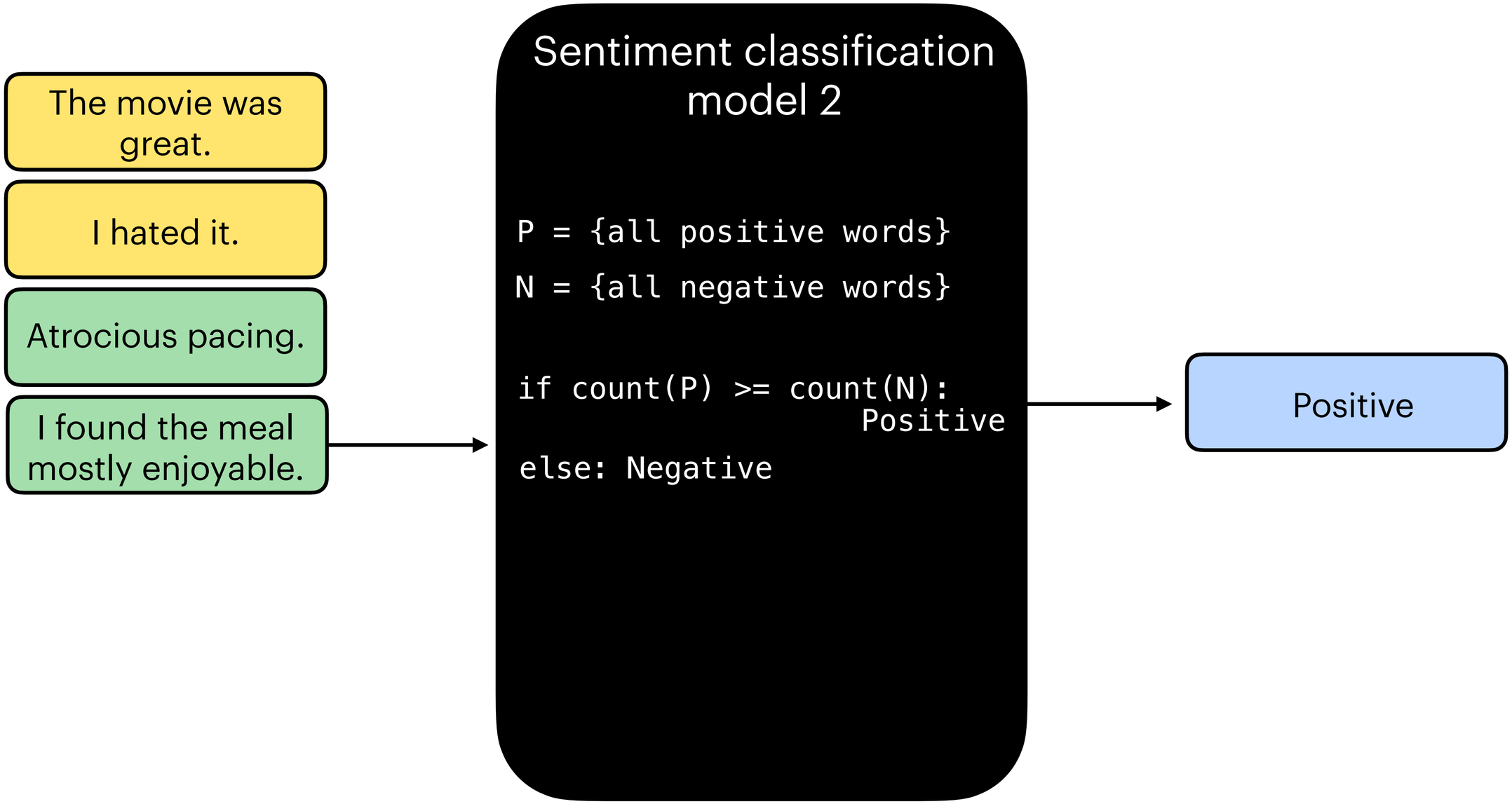
The Limitations of Behavioral Testing



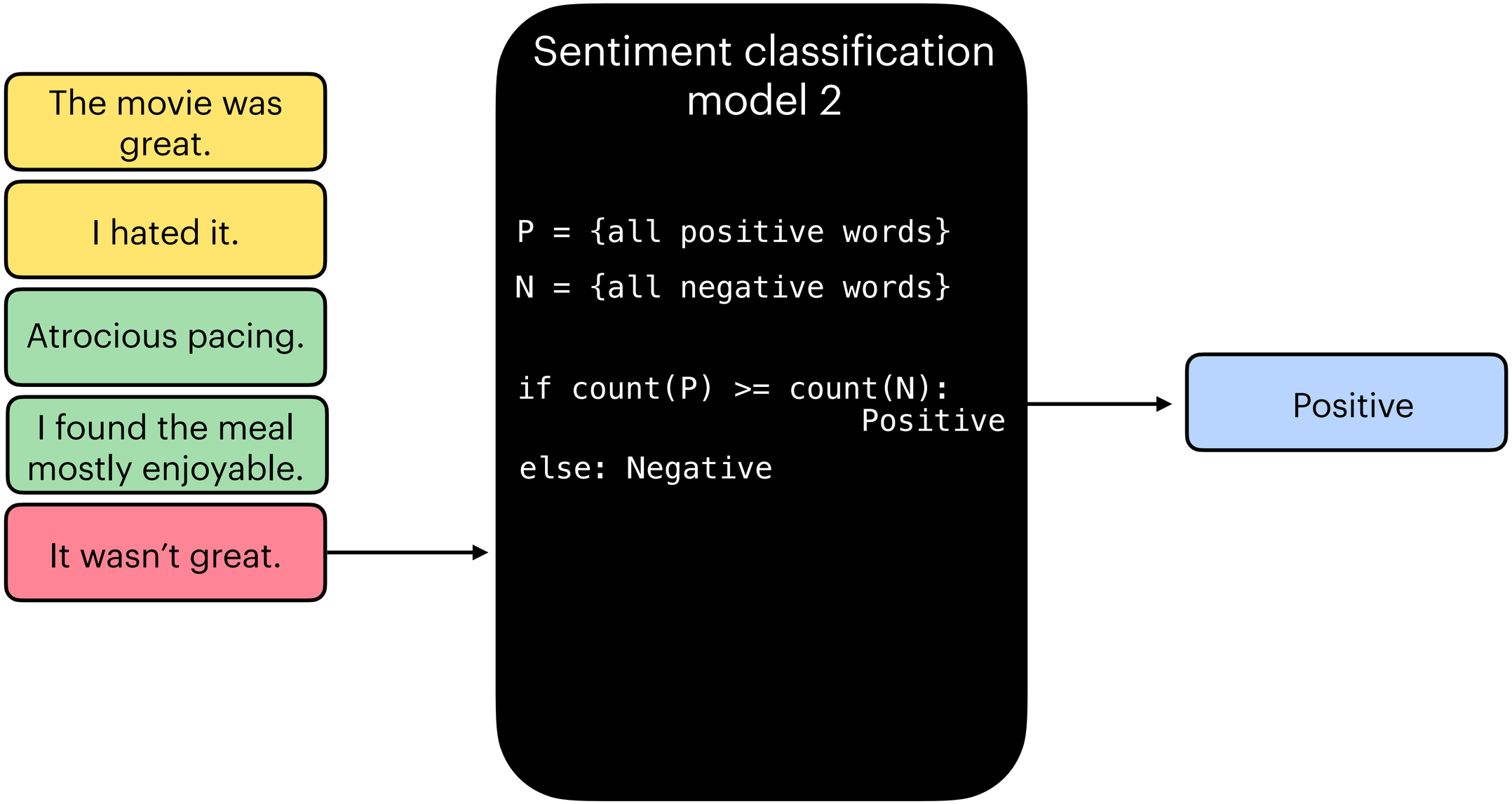
The Limitations of Behavioral Testing



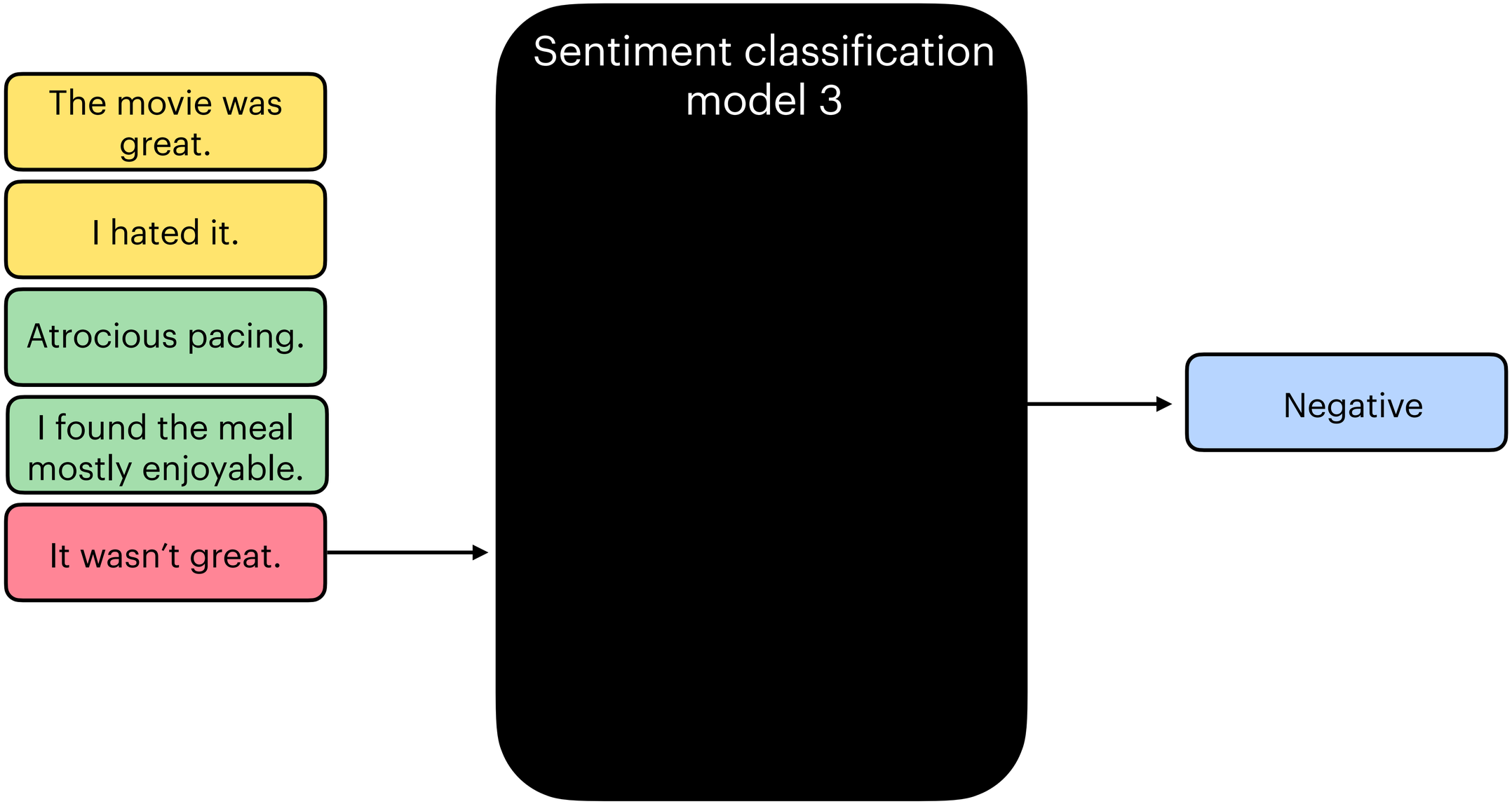
The Limitations of Behavioral Testing



The Limitations of Behavioral Testing



The Limitations of Behavioral Testing



Interpretability research asks ***why*** and ***how***?

Why did the model do that?

How is the model achieving these behaviors?

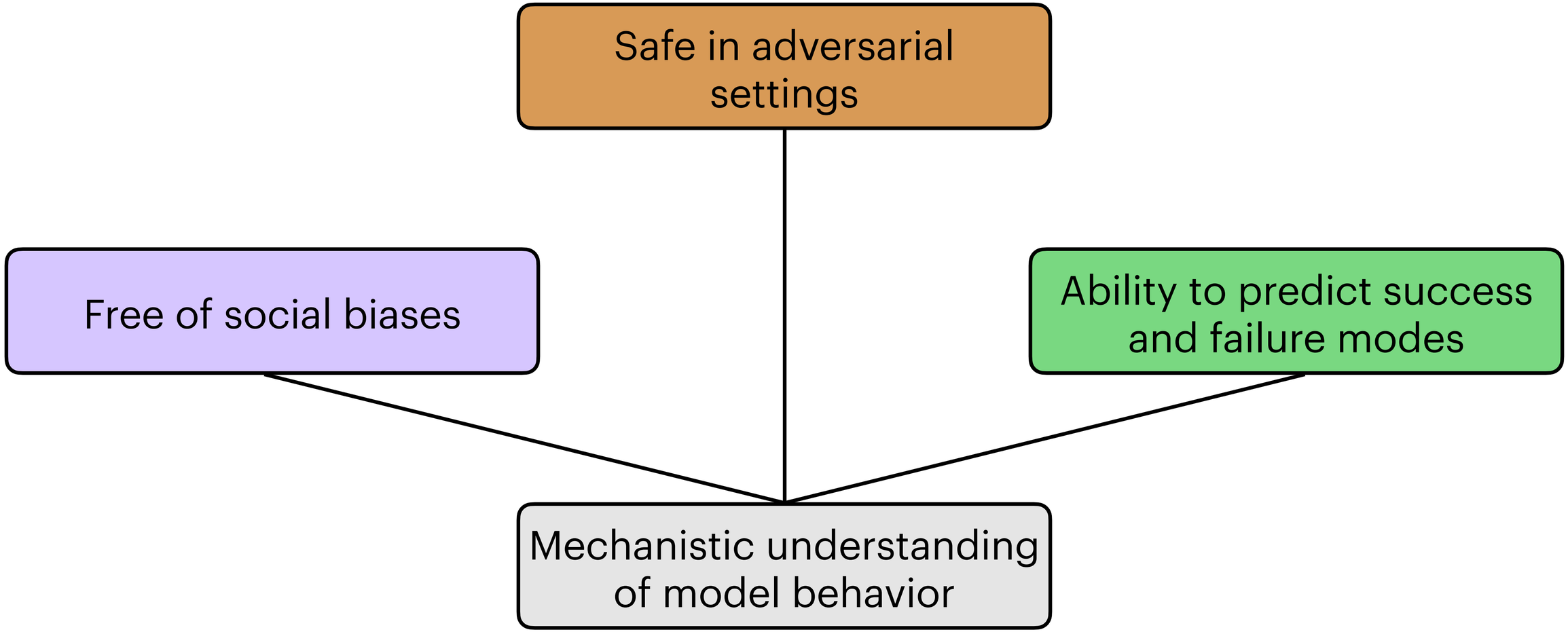
Critical Goals

Safe in adversarial settings

Free of social biases

Ability to predict success and failure modes

Critical Goals



Outline

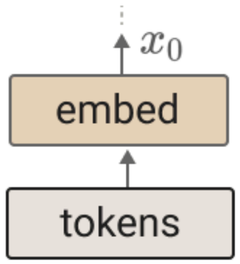
1. Probing

2. Localization and circuits

3. Applications: Debiasing and model editing

A Quick Review of Transformers

[Elhage et al., 2021]



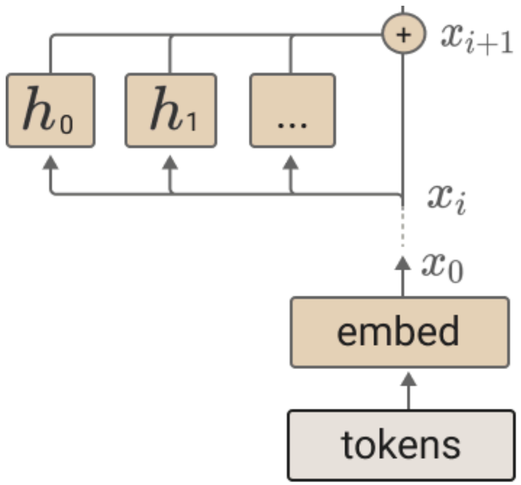
Token embedding.

$$x_0 = W_E t$$

Nelson Elhage et al. (2021). "A Mathematical Framework for Transformer Circuits." *Anthropic*.

A Quick Review of Transformers

[Elhage et al., 2021]



Each attention head, h, is run and added to the residual stream.

$$x_{i+1} = x_i + \sum_{h \in H_i} h(x_i)$$

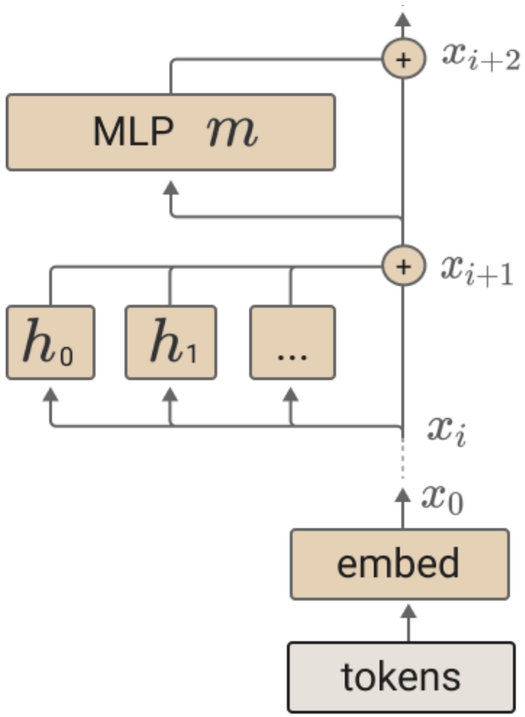
Token embedding.

$$x_0 = W_E t$$

Nelson Elhage et al. (2021). "A Mathematical Framework for Transformer Circuits." *Anthropic*.

A Quick Review of Transformers

[Elhage et al., 2021]



An MLP layer, m, is run and added to the residual stream.

$$x_{i+2} = x_{i+1} + m(x_{i+1})$$

Each attention head, h, is run and added to the residual stream.

$$x_{i+1} = x_i + \sum_{h \in H_i} h(x_i)$$

One residual block

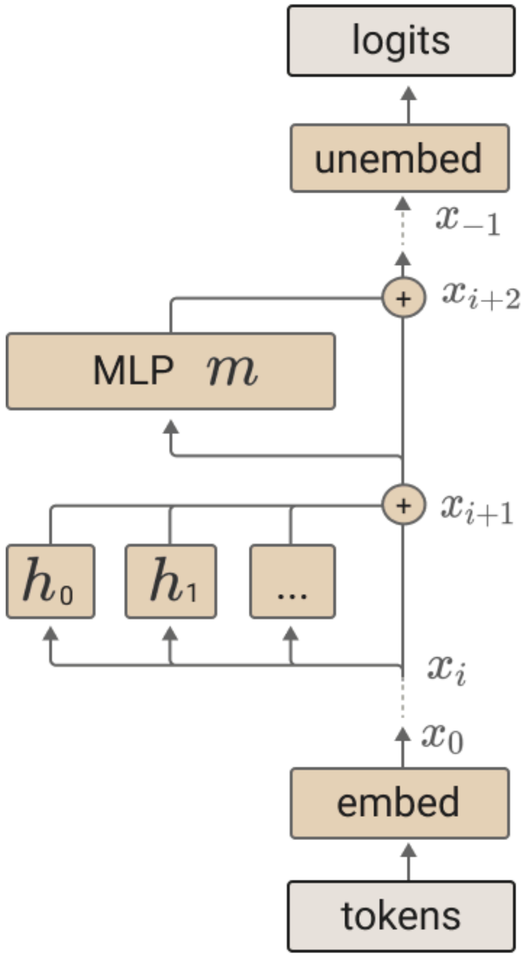
Token embedding.

$$x_0 = W_E t$$

Nelson Elhage et al. (2021). "A Mathematical Framework for Transformer Circuits." *Anthropic*.

A Quick Review of Transformers

[Elhage et al., 2021]



The final logits are produced by applying the unembedding.

$$T(t) = W_U x_{-1}$$

An MLP layer, m, is run and added to the residual stream.

$$x_{i+2} = x_{i+1} + m(x_{i+1})$$

Each attention head, h, is run and added to the residual stream.

$$x_{i+1} = x_i + \sum_{h \in H_i} h(x_i)$$

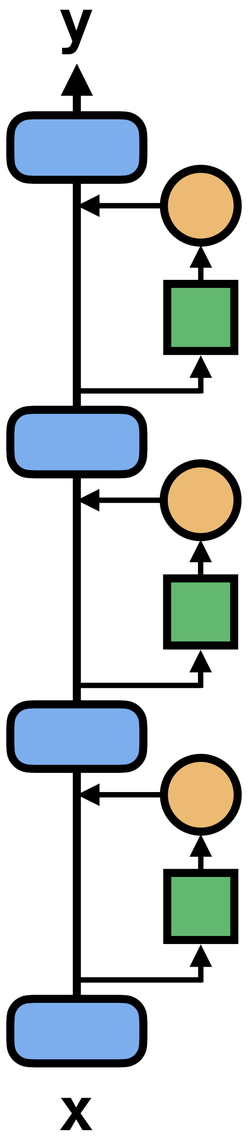
One residual block

Token embedding.

$$x_0 = W_E t$$

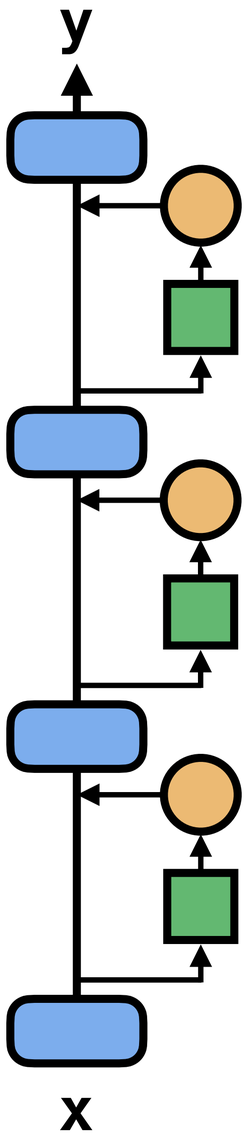
Nelson Elhage et al. (2021). "A Mathematical Framework for Transformer Circuits." *Anthropic*.

A Quick Review of Transformers



attention
MLP
residual vector

A Quick Review of Transformers



What concepts does this model represent?

attention
MLP
residual vector

Probing

- **Core idea**: use small supervised models (probes) to investigate what is encoded in the representations of larger models.
  - Popular from 2018 to 2021

- Can yield valuable insights, but we must be careful:
  - If the probe is too powerful, it can learn the task itself instead of revealing what's in the model.
  - Probes typically cannot tell us whether the information we find has any *causal* bearing on model behavior.

Probing

Method

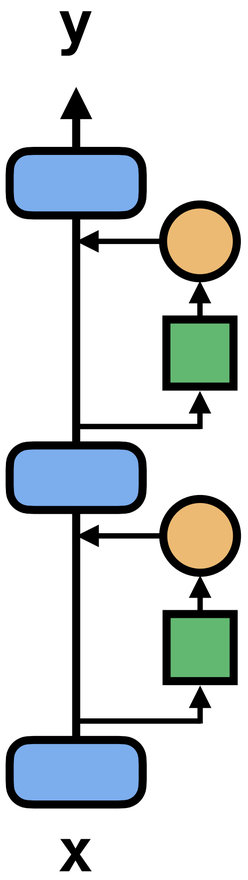
1. Hypothesize some feature, concept, or internal structure encoded by the model.

2. Design a supervised task to act as a proxy for the target structure.

3. Identify where in the model you believe the structure is encoded.

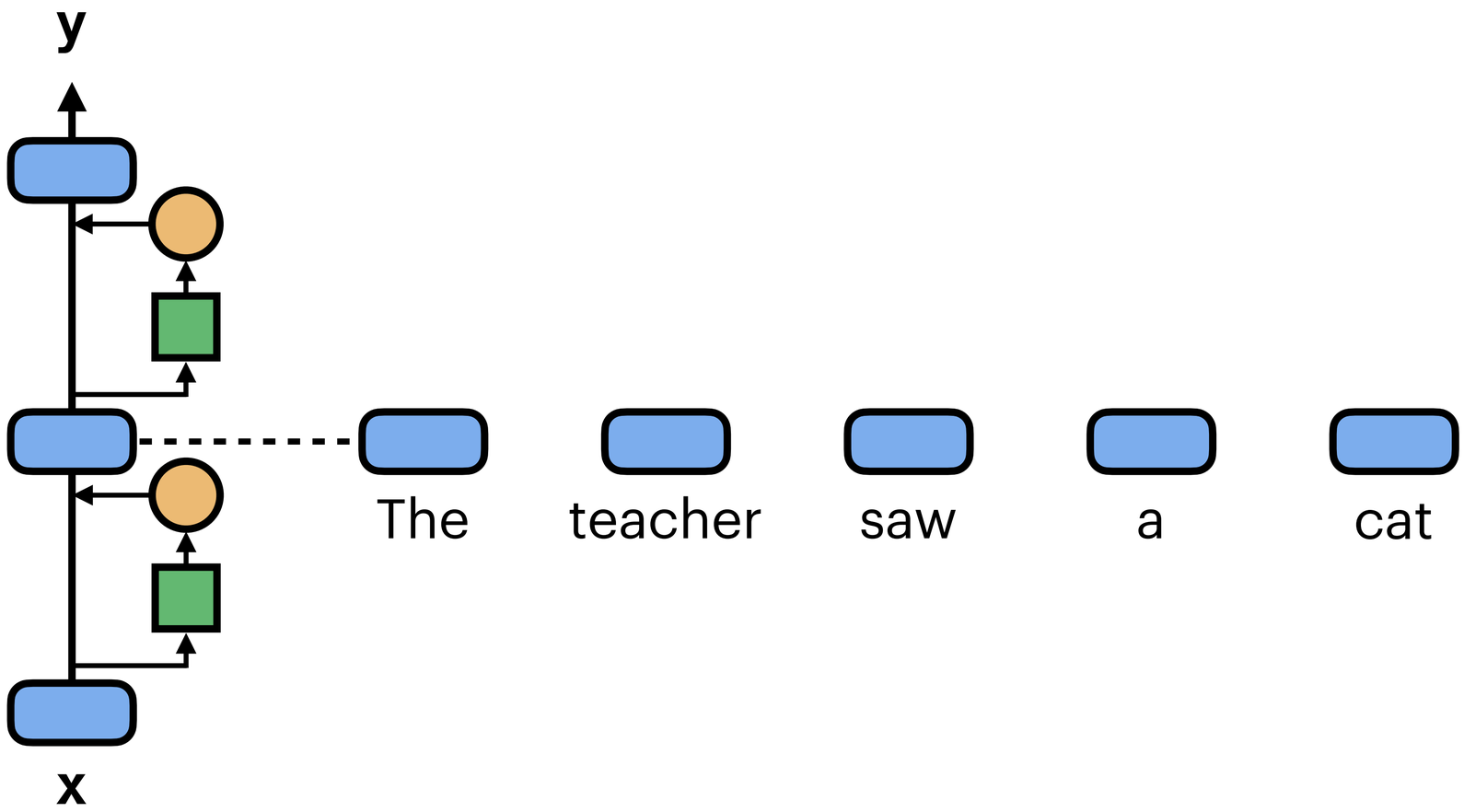
4. Train probes on the selected site(s) using the supervised task dataset.

Probing



Hypothesis:
Our model encodes part-of-speech information.

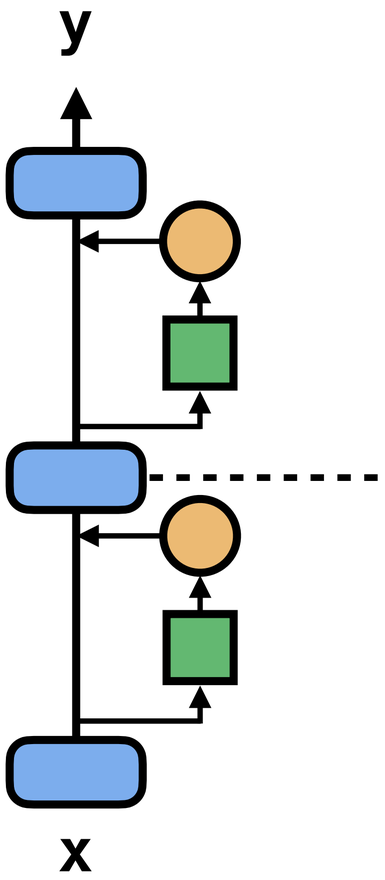
Probing



attention MLP
residual vector

Hypothesis:
The middle layer encodes part-of-speech information.

Probing



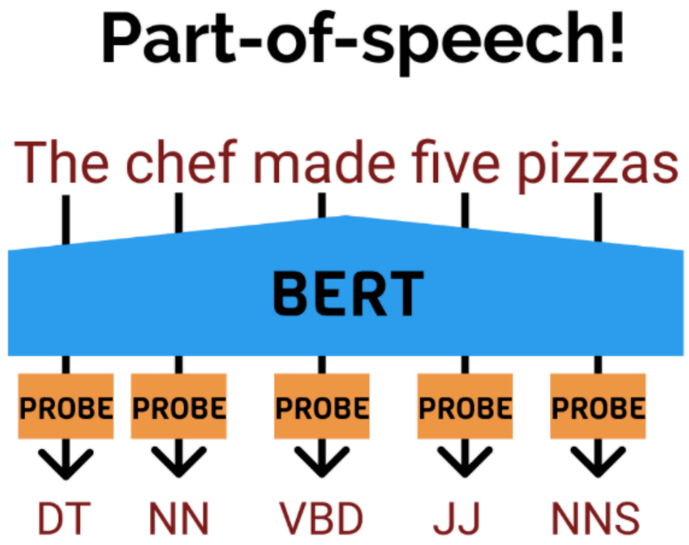
probe = SmallLinearClassifier(X, y)

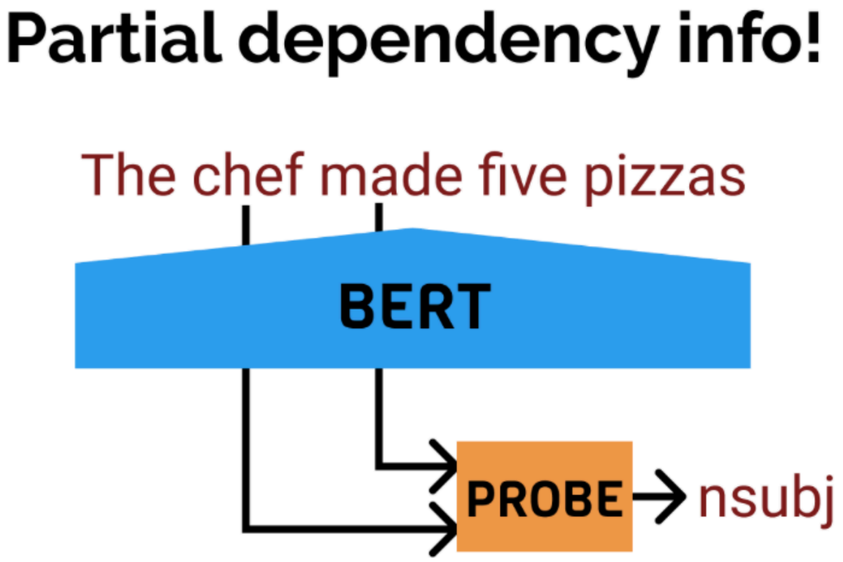
h_1	h_2	h_3	h_4	h_5
The	teacher	saw	a	cat
DET	NOUN	VERB	DET	NOUN

X	y
h_1	0
h_2	1
h_3	2
h_4	0
h_5	1

attention MLP
residual vector

Hypothesis:
The middle layer encodes part-of-speech information.





- If you can make it into a classification dataset, you can probe for it! Here are some other things people have probed models for:
  - Coreference **[Tenney et al., 2019]**
  - Full dependency parses **[Hewitt & Manning, 2019]**
  - Truthful/hallucinated content **[Marks & Tegmark, 2024; Obeso et al., 2025]**

Probing

Intuition

- Probes are supervised models that map from *frozen model representations* to *labels*.
- This is difficult to distinguish from simply fitting a supervised model as usual.
- At least some of the information we find is likely to be encoded in the probe itself.
- More powerful probes might “find” more information by simply storing this information in the probe parameters.

Selectivity

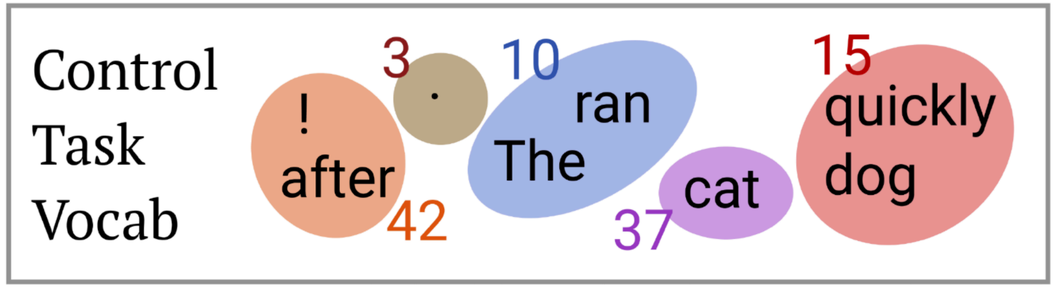
Control Task

A random task with the same input/label structure as the target task

- POS tagging: words with random fixed tags
- Dependency labeling: word pairs assigned random fixed dependency relations

Selectivity

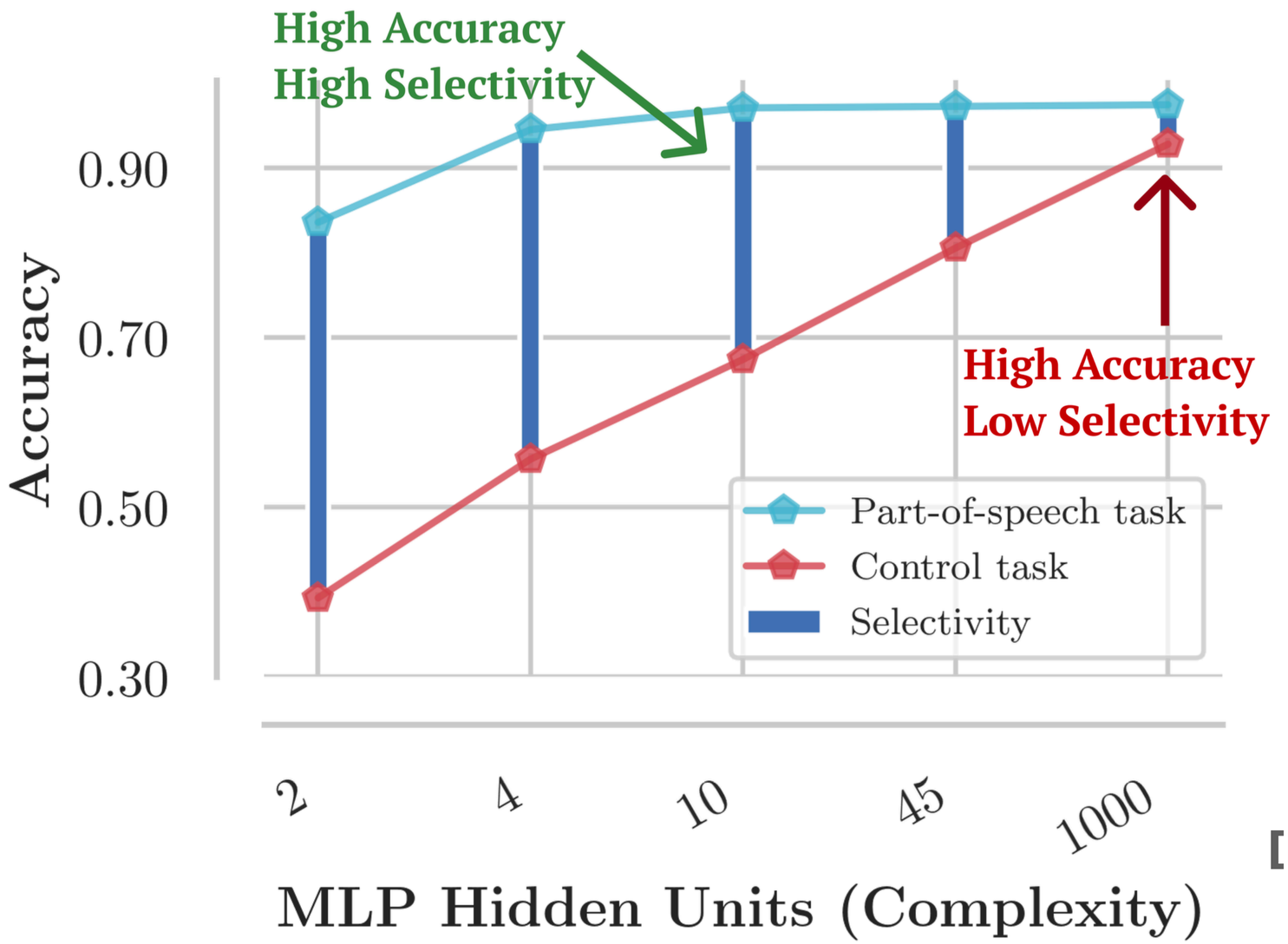
The difference between probe performance on the real task vs. the control task



Sentence 1	The	cat	ran	quickly	.
Part-of-speech	DT	NN	VBD	RB	.
Control task	10	37	10	15	3

Sentence 2	The	dog	ran	after	!
Part-of-speech	DT	NN	VBD	IN	.
Control task	10	15	10	42	42

Selectivity



[Hewitt & Liang, 2019]

John Hewitt & Percy Liang (2019). "Designing and Interpreting Probes with Control Tasks." *EMNLP*.

Unsupervised Probing

There are ways to characterize representations without labeled data!

- Inspect attention weights **[Clark et al., 2019; Manning et al., 2020]**
- Linearly transform hidden representations to identify latent syntactic structure **[Hewitt & Manning, 2019]**
- Singular Vector Canonical Correlation Analysis **[Saphra & Lopez, 2019]**
- Arguably: sparse autoencoders **[Cunningham et al. 2023; Bricken et al., 2023]**

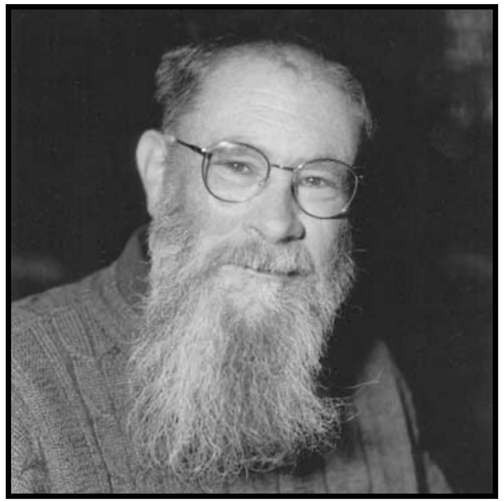
Mechanistic Interpretability

- Probing works by correlating parts of a model's representations to human concepts.

- **Mechanistic interpretability:** We'd like to understand how a model works by *understanding the causal roles of its internal computations*.
- How can we understand what a neuron does? An attention head?
- How do we know which components are important for a given task?
- **Interventions** do a lot of heavy lifting in this field.

Correlation vs. Causation

The **counterfactual theory of causality** holds the following:



"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."

—David Lewis

Correlations establish that two events co-occur often.

Causation posits a *counterfactual dependence* between the events.

Mechanistic Interpretability

A Case Study with Indirect Object Identification

Let's do some experiments to understand how models perform **indirect object identification** (IOI).

When **John** and **Mary** went to the store, **Mary** gave a book to ___

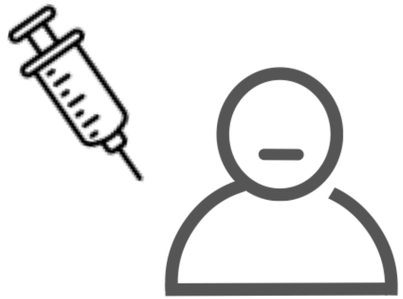
INDIRECT OBJECT — SUBJECT — SUBJECT

How does the model know which name to predict?

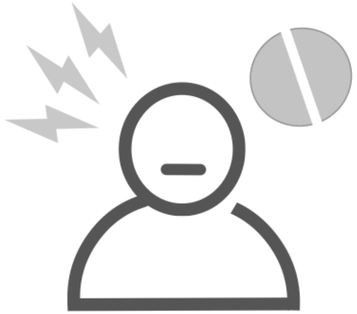
To find out, let's find the most causally relevant attention heads, and then inspect them.

Causal Mediation Analysis

Theory

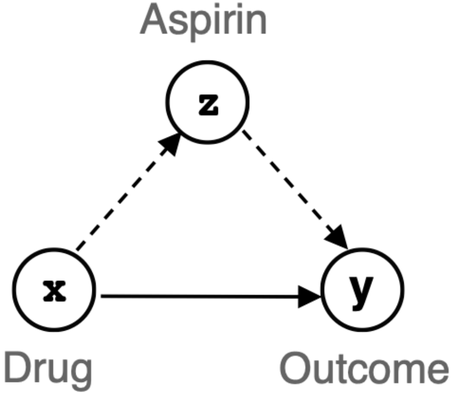


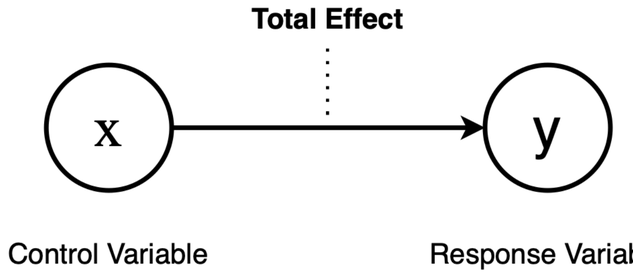
Administer drug

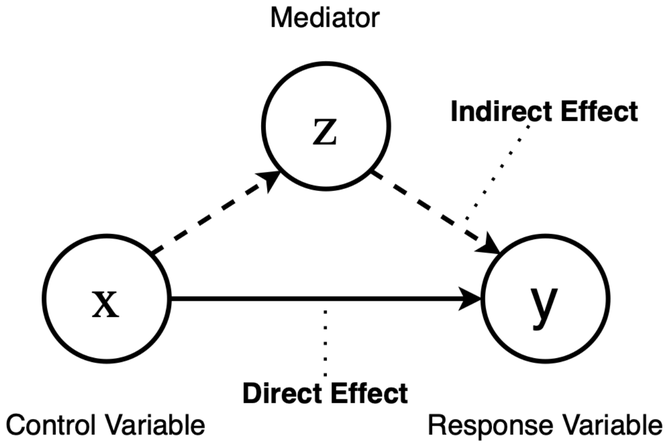


Drug causes headache, aspirin taking

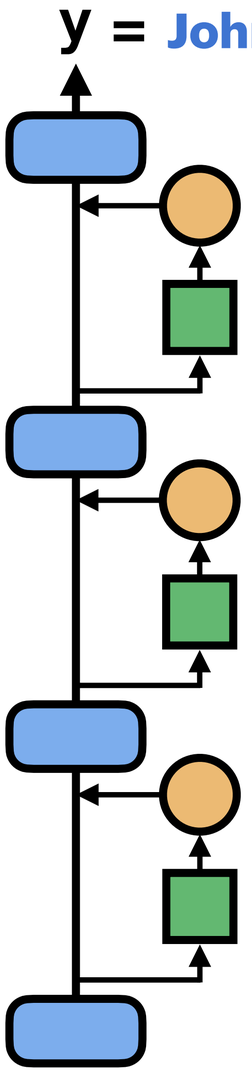
Drug + aspirin → health outcome

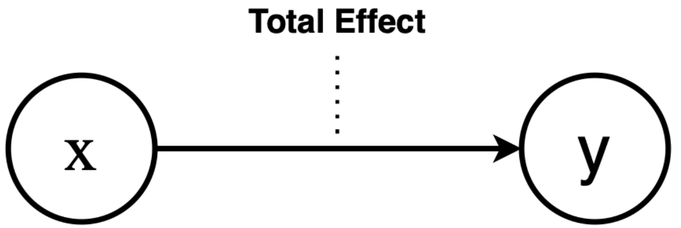






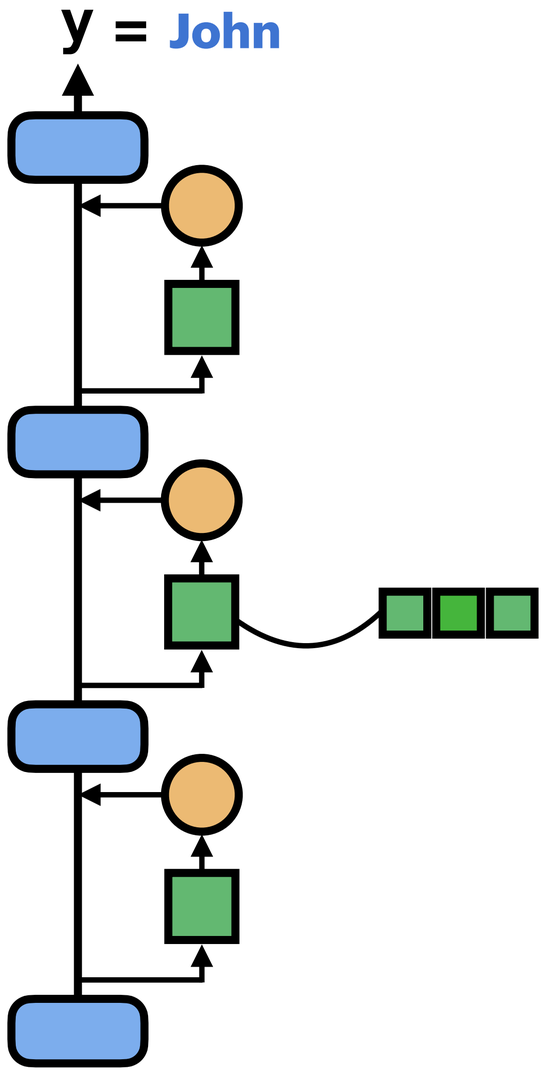
Causal Mediation Analysis





x = When **John** and **Mary** went to the store, **Mary** gave a book to ___

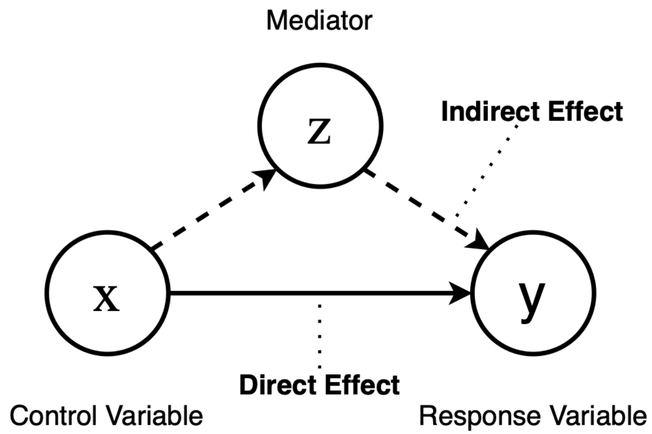
Causal Mediation Analysis

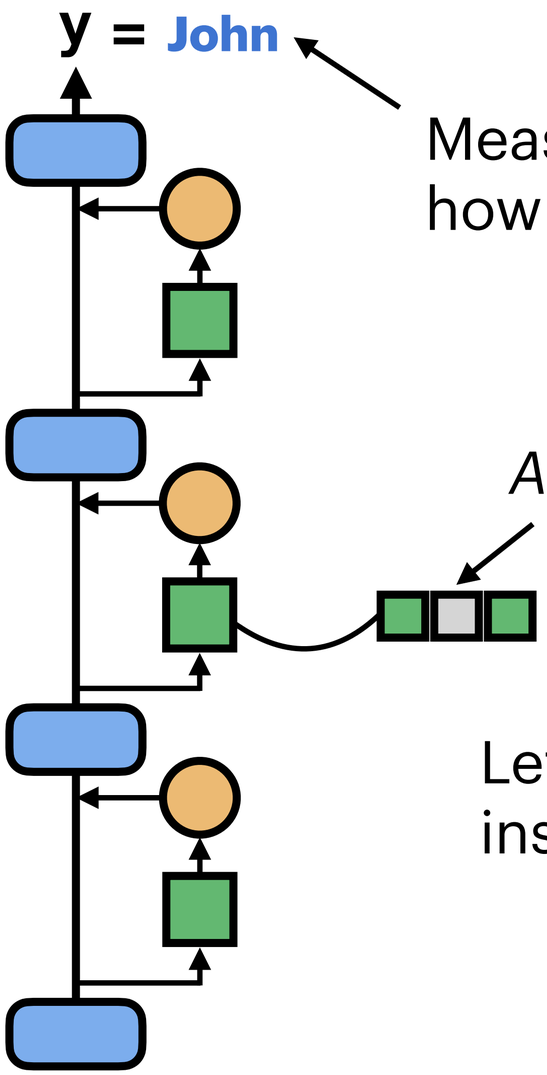


attention
MLP
residual vector

x = When **John** and **Mary** went to the store, **Mary** gave a book to ___

Causal Mediation Analysis





Measure **indirect effect**:
how much did the ablation change **y**?

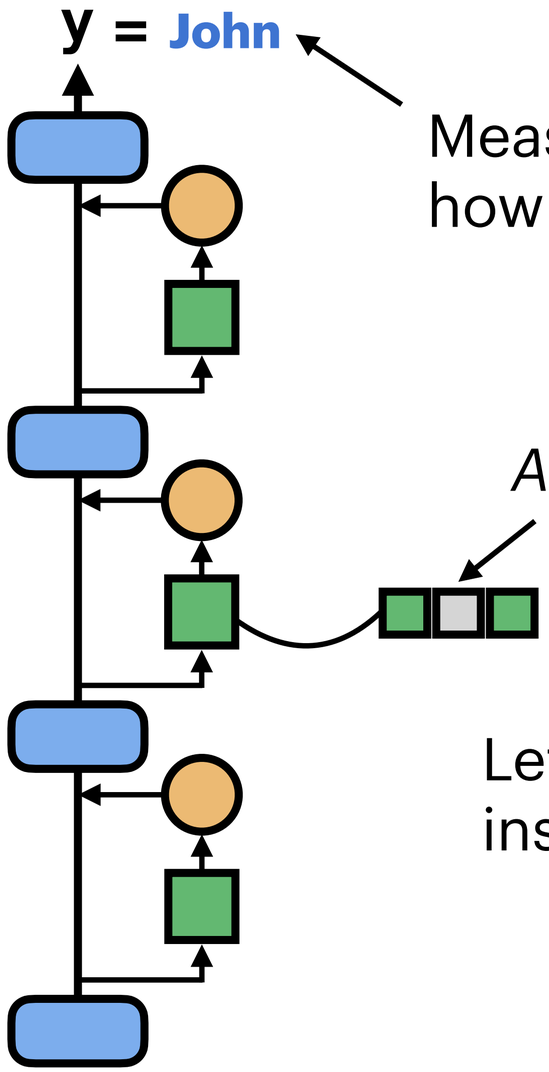
Ablate this head

Let's repeat this for each head, and then inspect the ones that change **y** the most.

attention MLP
residual vector

x = When **John** and **Mary** went to the store, **Mary** gave a book to ___

Causal Mediation Analysis



Measure **indirect effect**:
how much did the ablation change **y**?

Ablate this head: set it to 0?

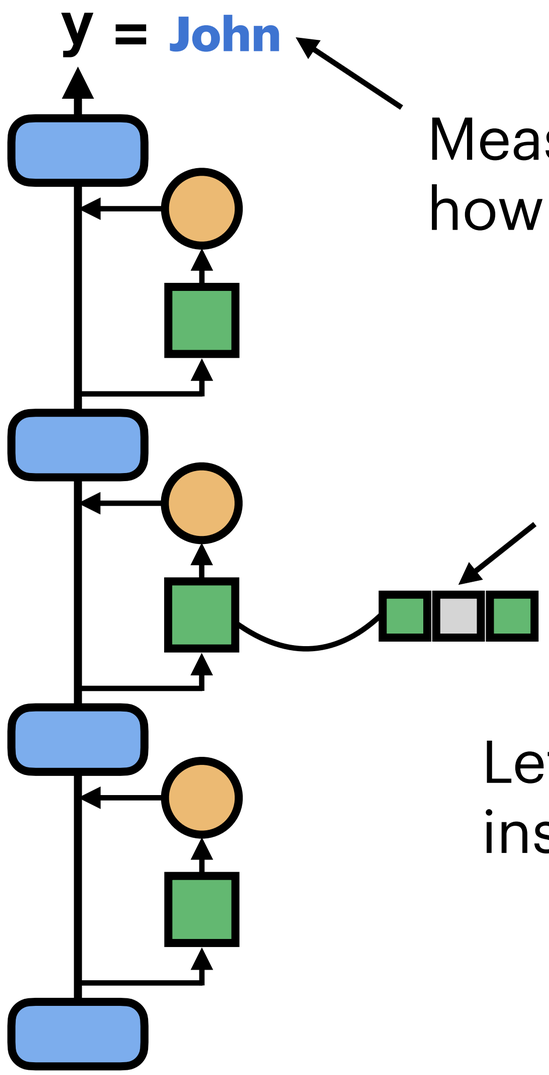
Problem: 0 doesn't have any inherent meaning

Let's repeat this for each head, and then inspect the ones that change **y** the most.

attention
MLP
residual vector

x = When **John** and **Mary** went to the store, **Mary** gave a book to ___

Causal Mediation Analysis



Measure **indirect effect**:
how much did the ablation change **y**?

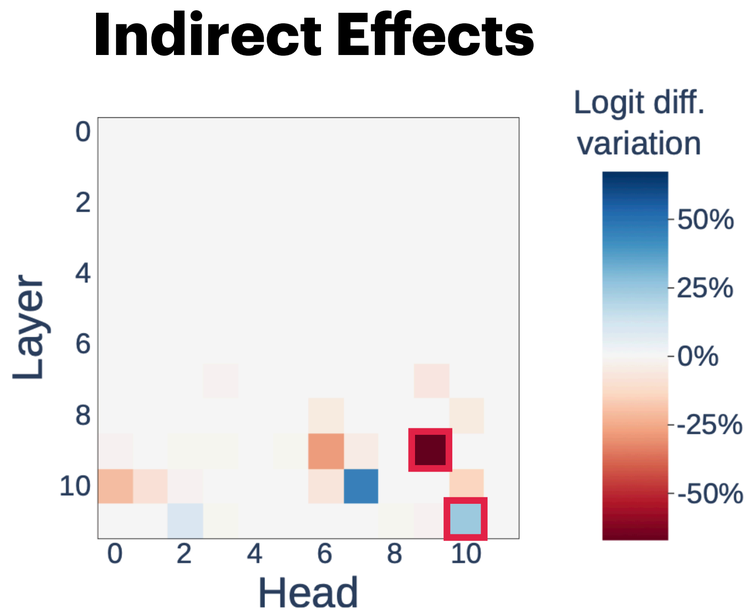
Ablate this head: set it to its *mean value* on some dataset

Let's repeat this for each head, and then inspect the ones that change **y** the most.

attention
MLP
residual vector

x = When **John** and **Mary** went to the store, **Mary** gave a book to ___

[Wang et al., 2023]

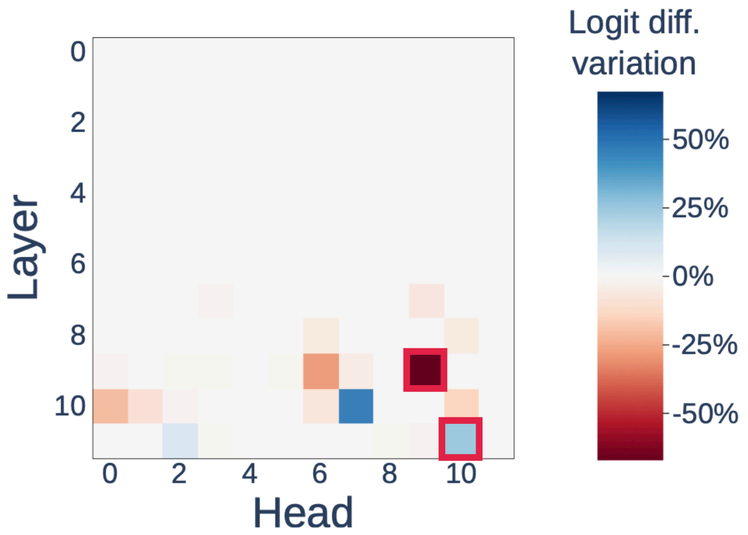


Kevin Wang et al. (2023). “Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 Small.” *ICLR*.

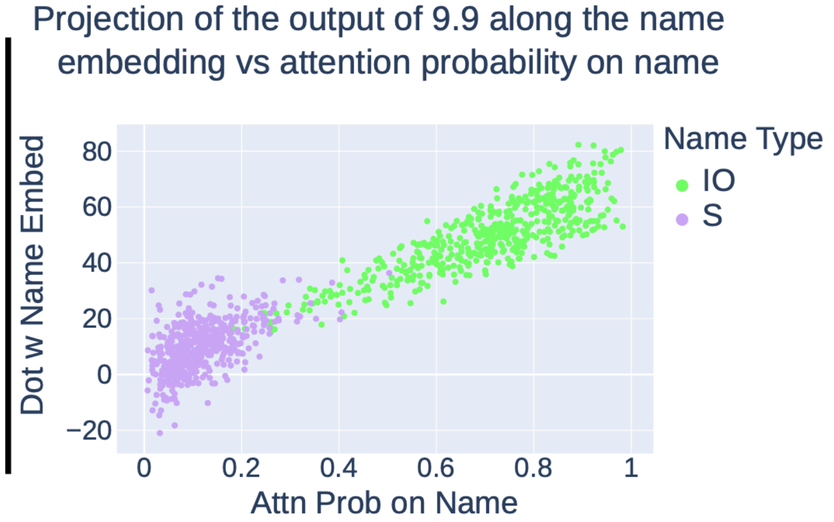
These heads are moving the name information from where it was first mentioned to just before where it's needed for prediction!

Name mover heads

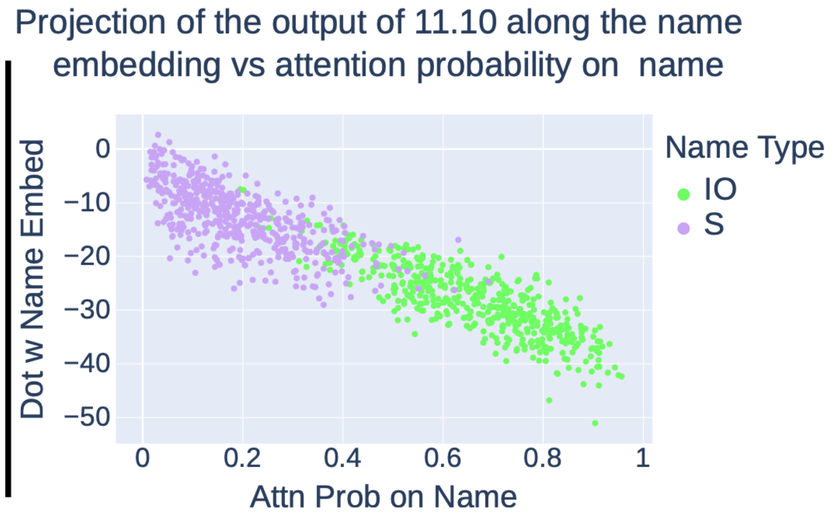
Indirect Effect



Name Mover Heads

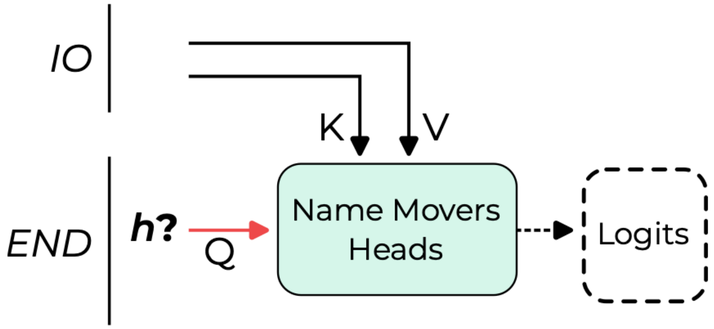


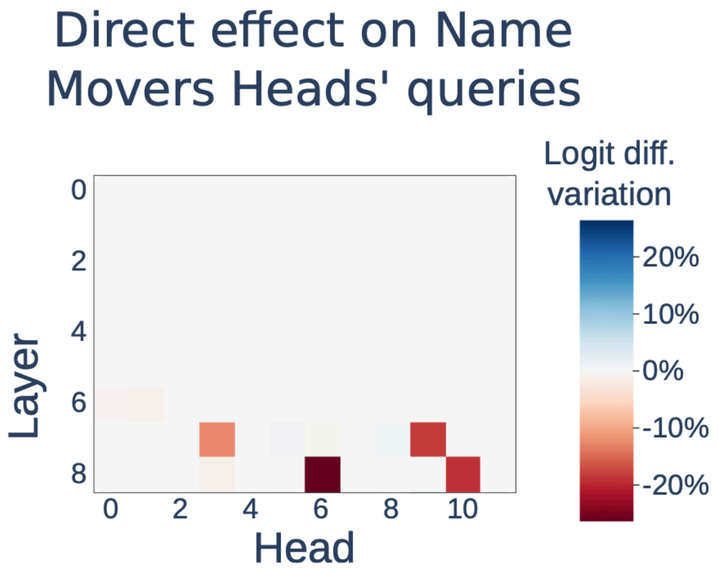
Negative Name Mover Heads

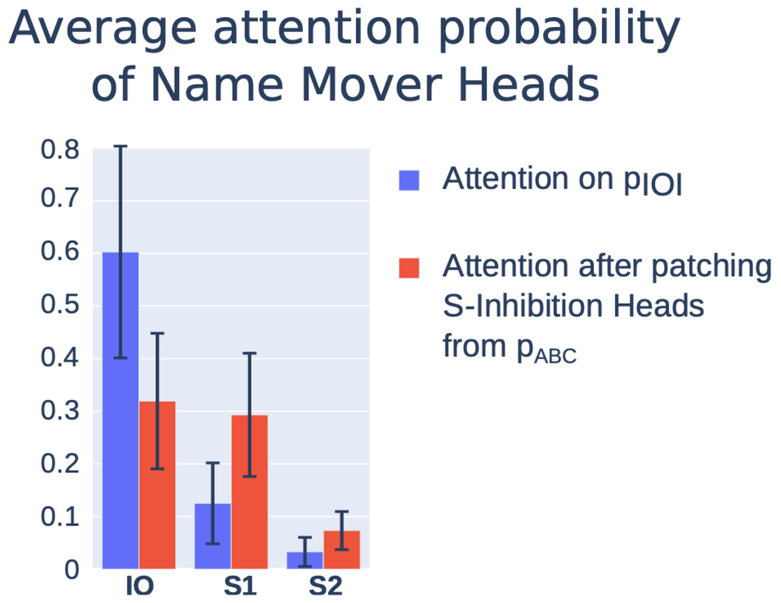


Kevin Wang et al. (2023). "Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 Small." *ICLR*.

[Wang et al., 2023]





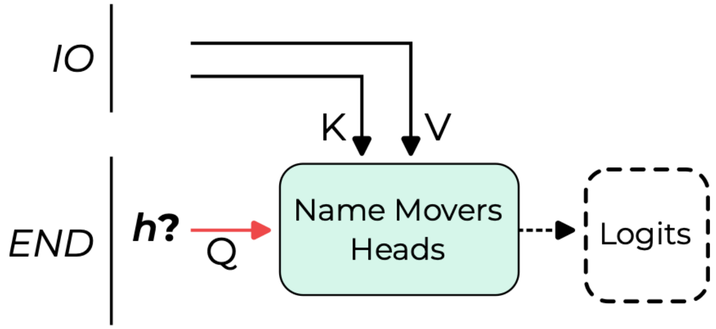


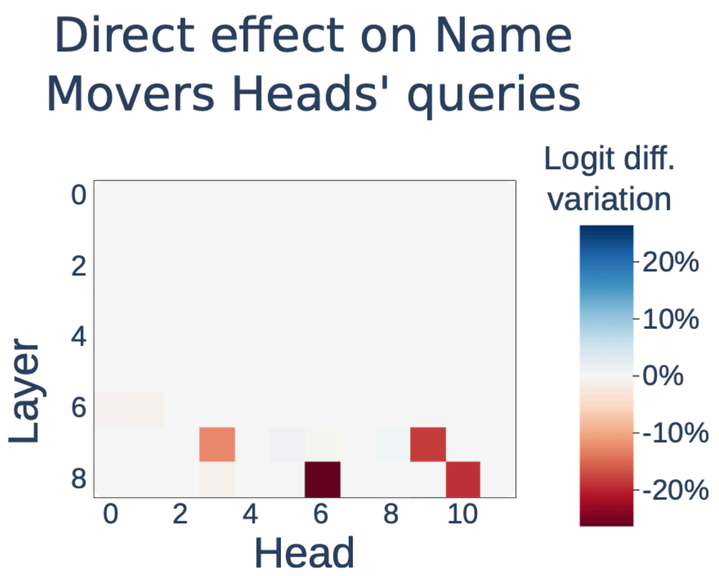
Kevin Wang et al. (2023). "Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 Small." *ICLR*.

[Wang et al., 2023]

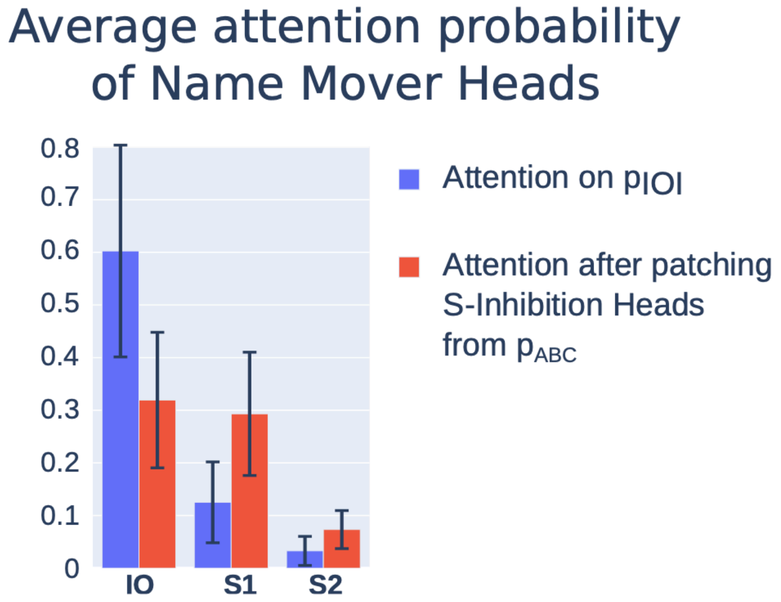
Patching them increases *p*(SUBJECT),
so their role is to decrease *p*(SUBJECT)

These heads directly affect
the name mover heads we found!





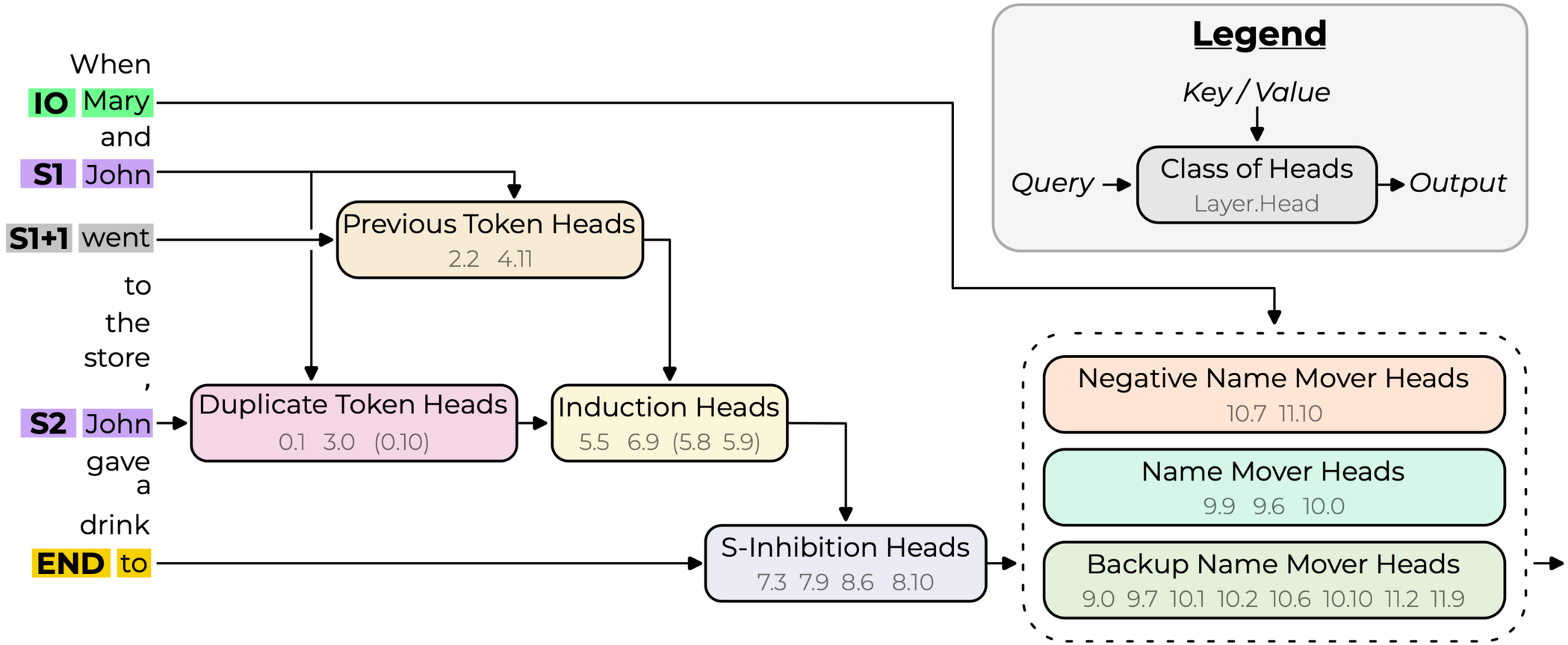
S-inhibition heads



Kevin Wang et al. (2023). "Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 Small." *ICLR*.

[Wang et al., 2023]

If we keep repeating this process until we find all the important heads...



We end up with a **circuit.**

Kevin Wang et al. (2023). "Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 Small." *ICLR*.

Circuits

Pros:

Enable deep understanding of a model behavior

Highly precise and causally efficacious

*Could allow us to remove undesirable computations?

Cons:

Requires a lot of human effort to find and understand

It’s not always easy to understand what model components do

*Can be automated, but is slow

Circuits

Pros:

Enable deep understanding of a model behavior

Highly precise and causally efficacious

***Could allow us to remove undesirable computations?**

Cons:

Requires a lot of human effort to find and understand

It’s not always easy to understand what model components do

***Can be automated, but is slow**

Can we use interpretability to fix models?

- Understanding how models work is nice...
- ...but ideally, we'd like to be able to *fix* models.
- We often don't know ahead of time what biases/features to look for.
- Retraining from scratch is extremely expensive (>$1M).
    - May not help remove biases
- Recent post-training methods may not be precise or effective enough.

Model Steering and Editing

Can we precisely **localize** where a behavior is being computed in a neural network?

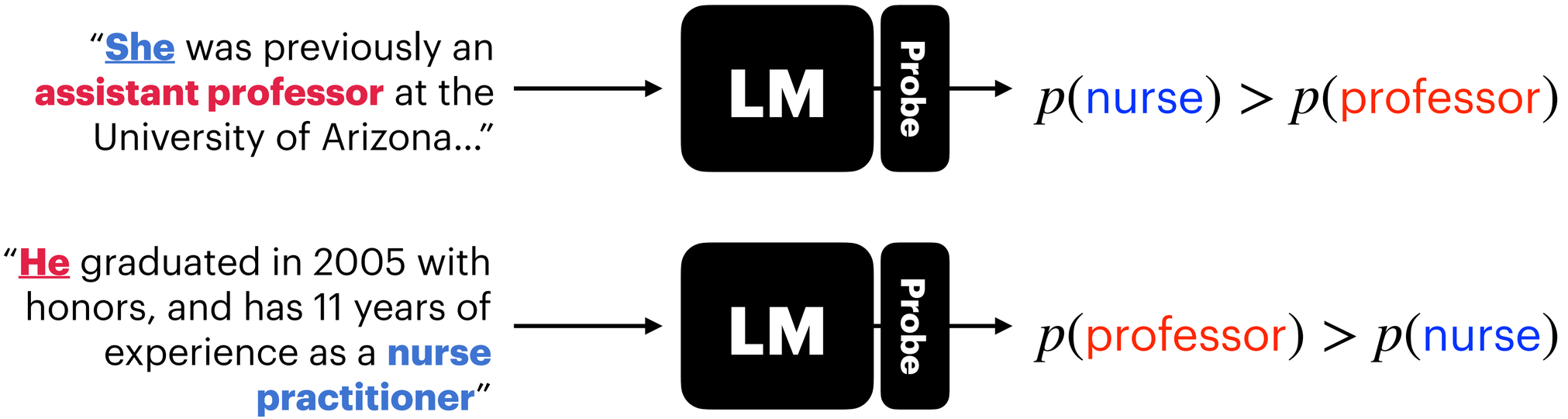
Can we perform fine-grained **interventions** to modify the mechanism or information underlying the behavior?

Yes! We'll go over two methods:

1. *Feature steering* with sparse feature circuits (SHIFT)
2. *Model editing* with ROME

Debiasing with Interpretability Tools

A Motivating Example



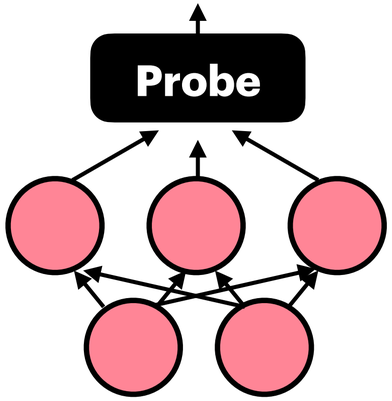
Let's *locate* and *remove* the components responsible for gender bias.

Causal Mediation Analysis

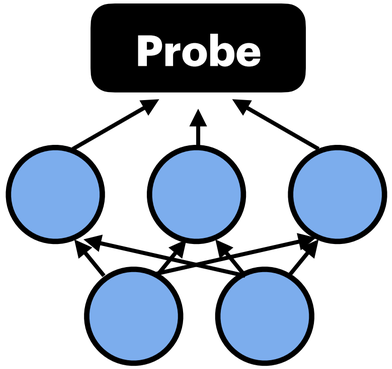
Method

1. Cache activations $\mathbf{z}$ and metric m given two minimally different inputs.

$$m = p(\text{nurse}) - p(\text{professor})$$



$x =$ They are a professor at...



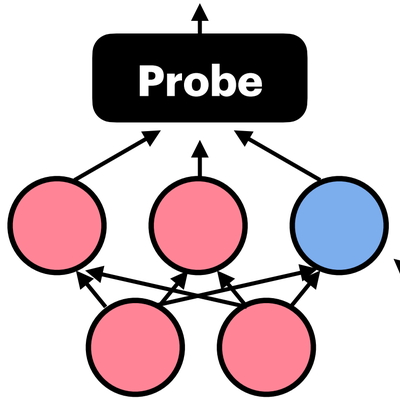
$x' =$ With a long career in nursing, ...

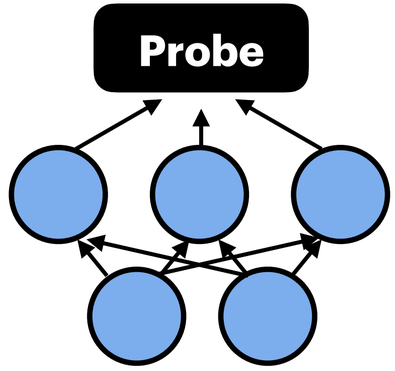
Causal Mediation Analysis

Method

1. Cache activations $\mathbf{z}$ and metric m given two minimally different inputs.

$$m = p(\text{nurse}) - p(\text{professor})$$





x = They are a professor at...

x' = With a long career in nursing, ...

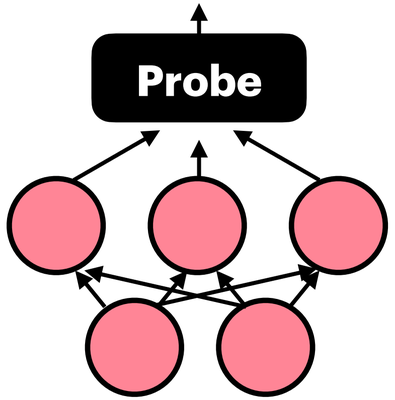
2. Perform counterfactual intervention to a neuron z, measure how this affects m.

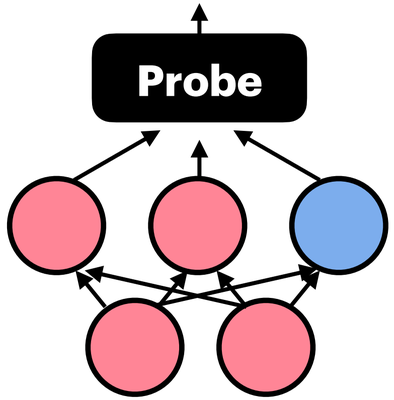
`swap-profession:`
Replace with activation from identical input where expected output is different.

Causal Mediation Analysis

Method

$m = p(\text{nurse}) - p(\text{professor})$ — Indirect effect → $p(\text{nurse}) - p(\text{professor})$





$x =$ They are a professor at...

$x =$ They are a professor at...

Indirect effect (IE): a *causal* measure of how much an intermediate variable influences the final output.

$$\text{IE}(m, x, x', z) = m(x \mid \text{do}(z = z(x'))) - m(x)$$

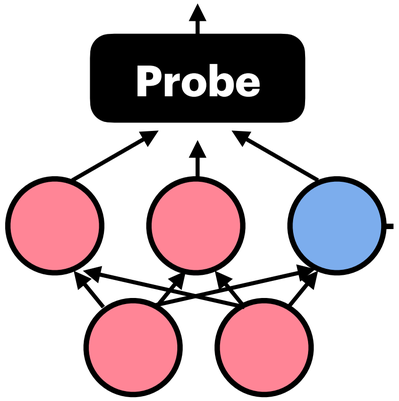
Causal Mediation Analysis

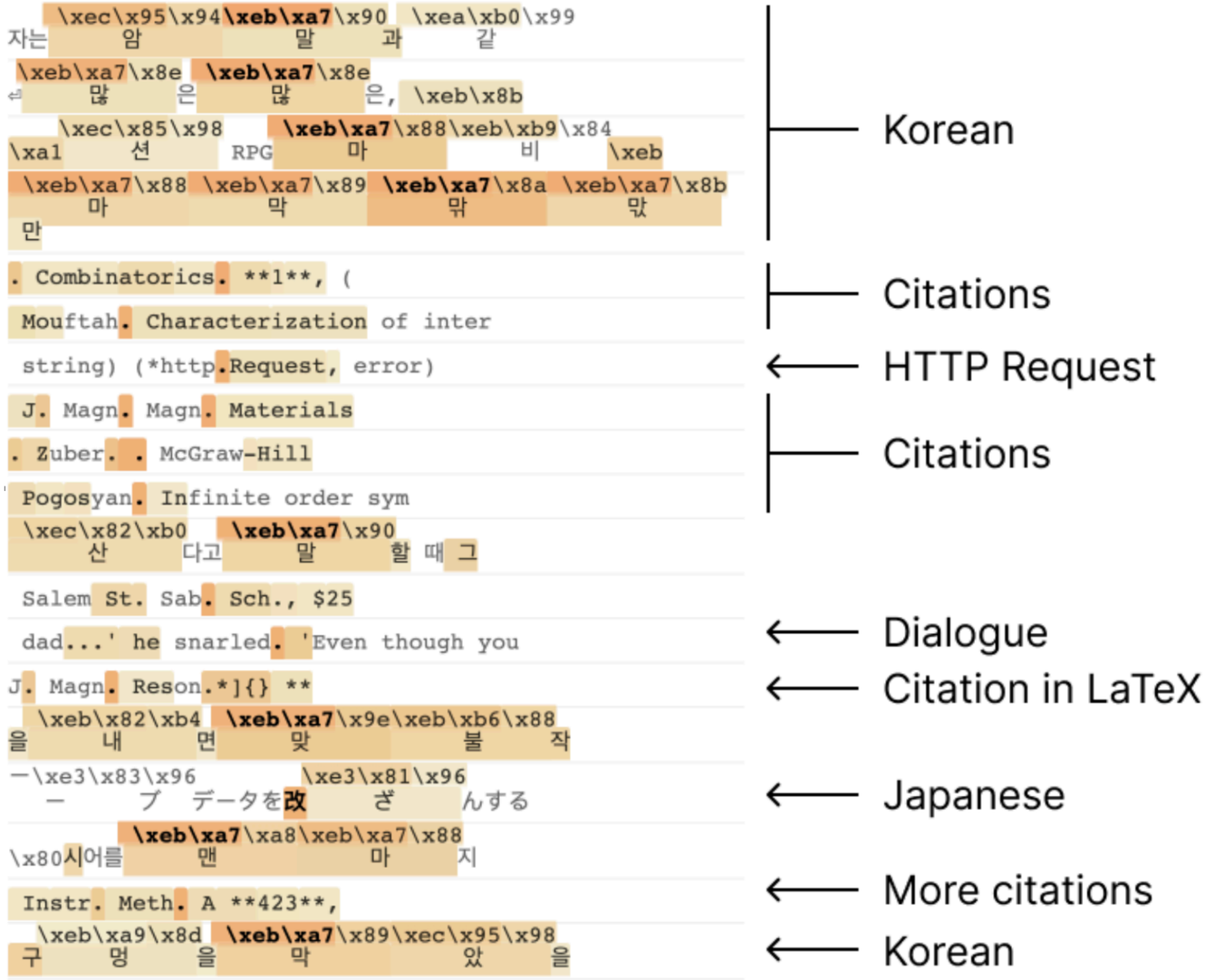
Removing Bias

- Let’s compute the IE for all neurons in the model.
- Then, let’s rank them by IE, and **ablate** the top k% to remove bias.
- **Problem:** representations are distributed!

Neurons are often polysemous.

$$m = p(\text{nurse}) - p(\text{professor})$$





Trenton Bricken et al. (2023). "Towards monosemanticity: Decomposing language models with dictionary learning." *Anthropic*.

[Bricken et al., 2023]

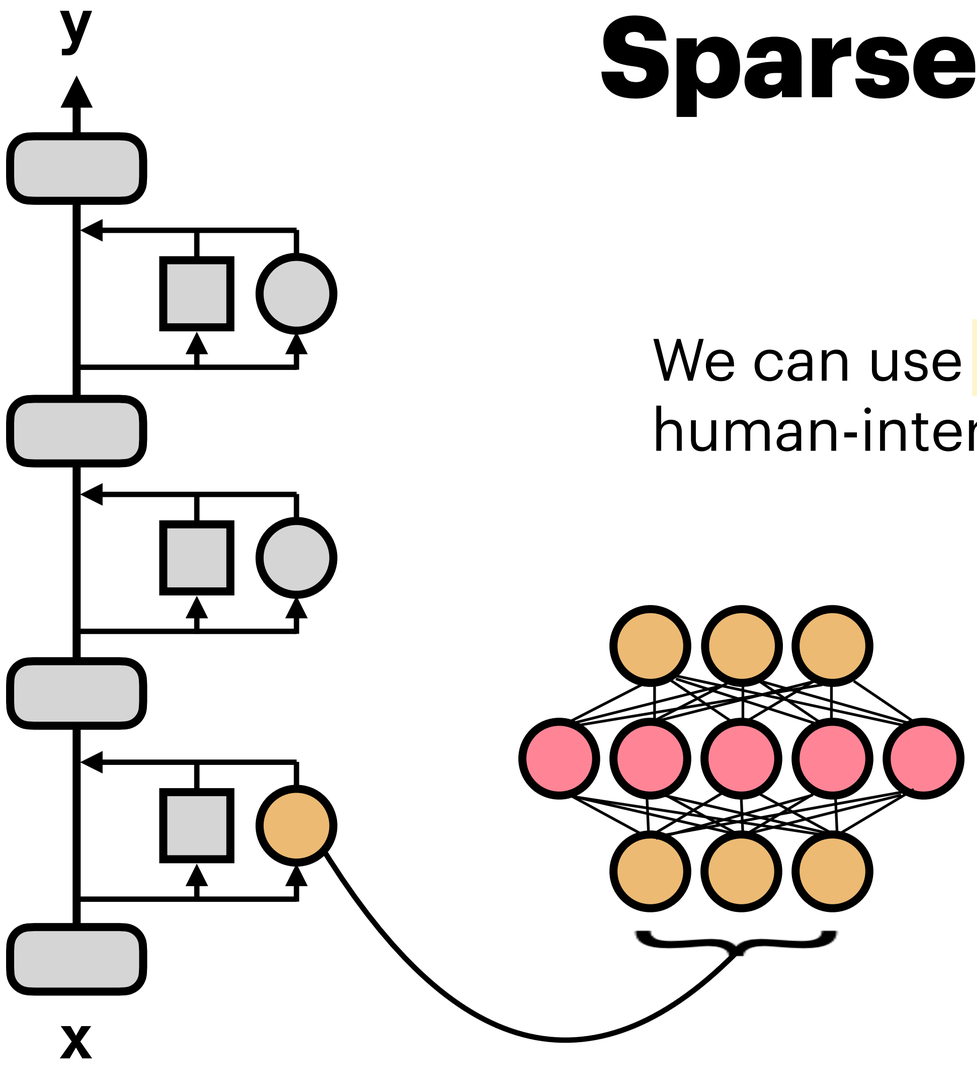
A Way Forward: Featurize!

- Convert dense neuron representations into **human-interpretable features** before we use causal mediation analysis.
- There are many ways to do this:
  - Use precisely controlled pairs of examples to learn rotations that isolate meaningful neurons **[Geiger et al., 2021]** or subspaces **[Geiger et al., 2023; Wu et al., 2024]**.
  - Use the direction learned by a probe **[Marks et al., 2024]**.
  - Use sparse autoencoders **[Cunningham et al., 2023; Bricken et al. 2023]**.

Requires us to know in advance what we're looking for.

Allows us to locate causes we don't anticipate!

Sparse Features



We can use **sparse autoencoders** (SAEs) to disentangle human-interpretable **features** from model components

$$\hat{\mathbf{x}} = W_d\mathbf{f} + \mathbf{b}_d$$

$$\mathbf{f} = \text{ReLU}(W_e(\mathbf{x} - \mathbf{b}_d) + \mathbf{b}_e)$$

$$\mathbf{x}$$

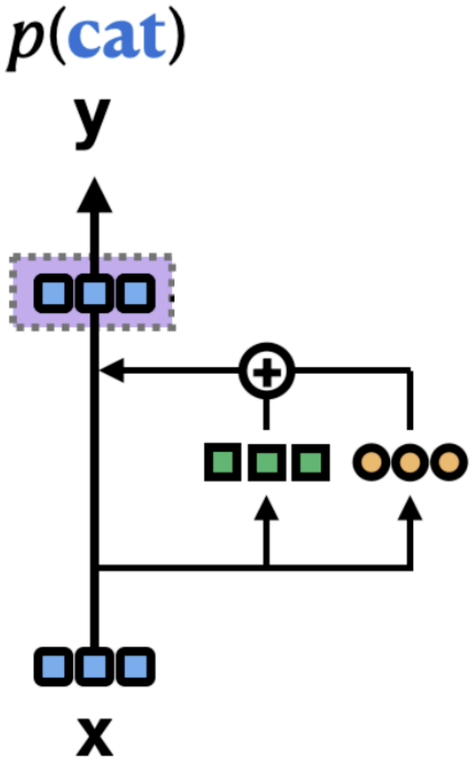
$$L = \sqrt{\text{MSE}(\mathbf{x}, \hat{\mathbf{x}})} + \lambda\|\mathbf{f}\|_1$$

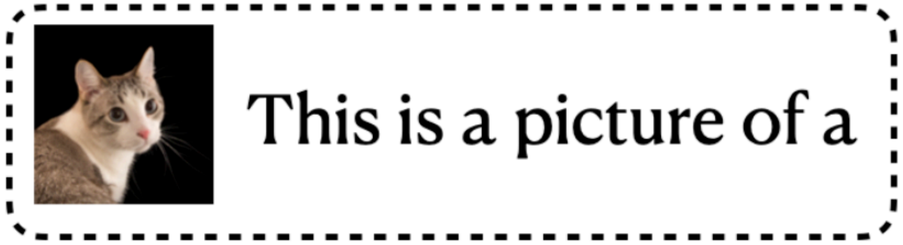
$$\mathbf{x} = \hat{\mathbf{x}} + \epsilon$$

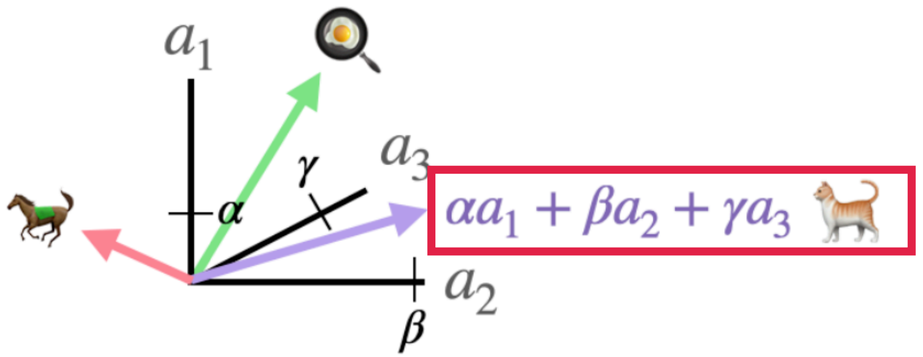
attention
MLP
vector state
SAE

Sparse Features

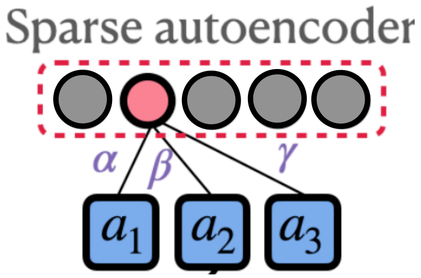
Intuition

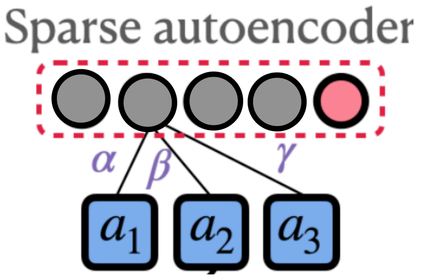






Examples





obau, the daughter of Ratu Sir George

office by a homeless woman named Lois Lang.

Benedict debate. But she has some thoughts on

of these creative↵women, the reader gets

↵↵"Ma'am?"↵↵"You

the physician who examined her body was unable to

you hear↵ her towards the end what

↵Norma and Sherryl suggest that there was

Words related to women

goal of our research program on innate immune sensors

4.↵↵His research interests include bioinformatics

.K.'s group are funded by the

Dave Lovinger?s Laboratory, investigates the

a Hungarian mathematician who works as a professor at

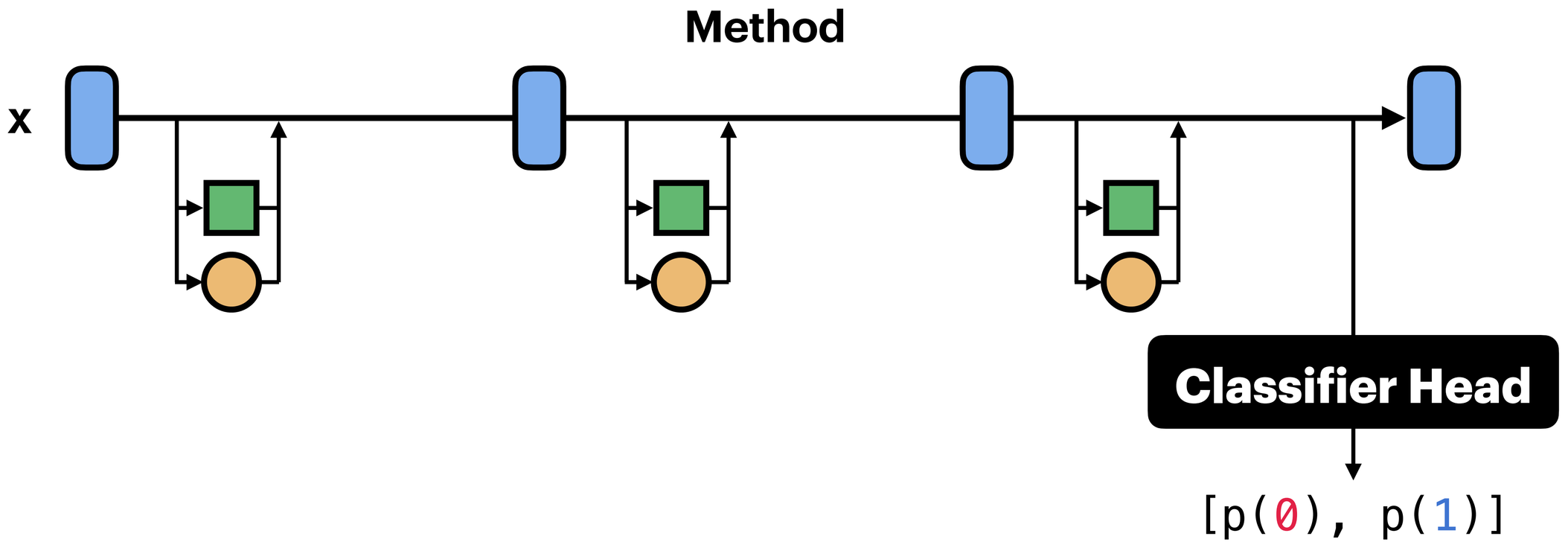
in the Kalluri laboratory, where both tumor

the Human Cognitive Neuroscience Unit at Northumbria University

Murakami's research team, which received a

Passages related to academia, research

SHIFT

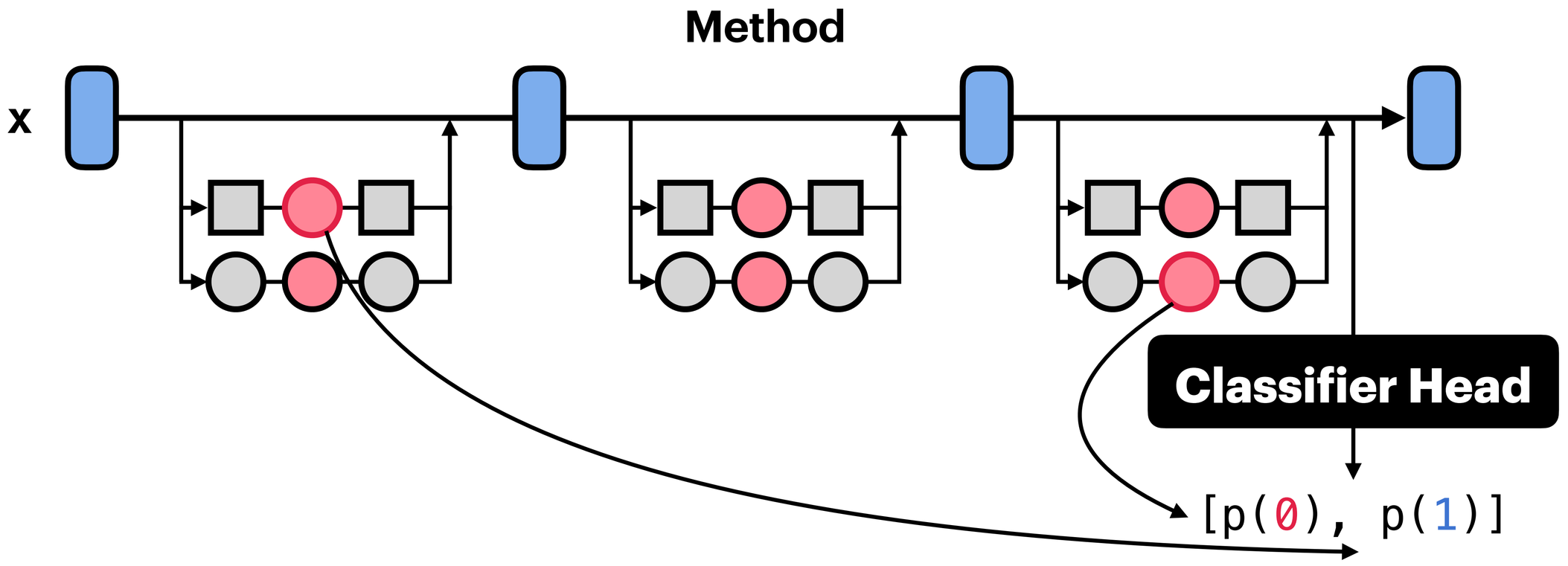


Task: classify profession

Acc.:

Profession : 63%
Gender: 87%

SHIFT



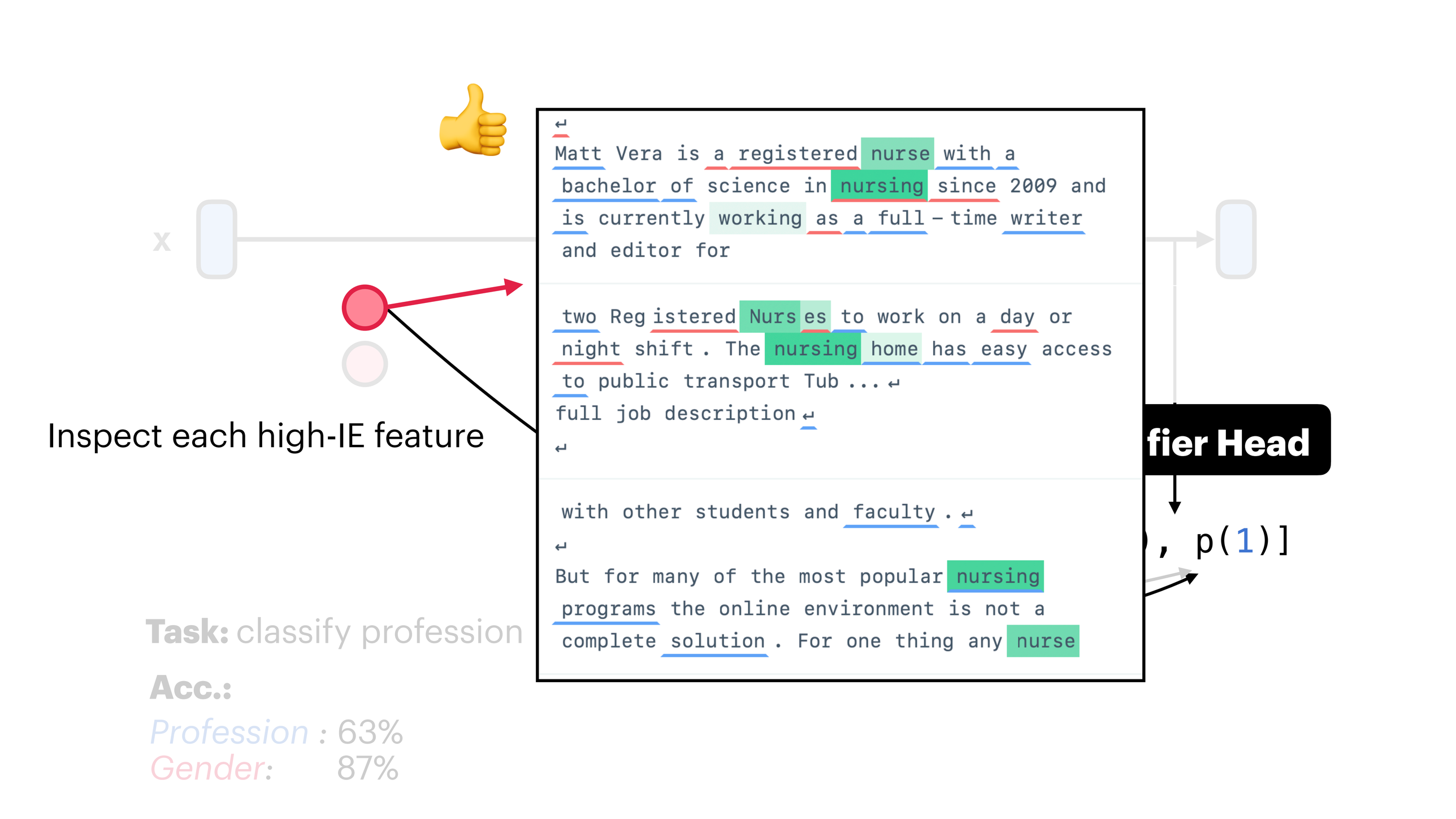
Task: classify profession

Acc.:

Profession : 63%

Gender: 87%

Look for features with high IE on classifier logits


Inspect each high-IE feature
x
↵
Matt Vera is a registered nurse with a
bachelor of science in nursing since 2009 and
is currently working as a full - time writer
and editor for
two Registered Nurses to work on a day or
night shift . The nursing home has easy access
to public transport Tub ... ↵
full job description ↵
↵
with other students and faculty . ↵
↵
But for many of the most popular nursing
programs the online environment is not a
complete solution . For one thing any nurse
fier Head
), p(1)]
Task: classify profession
Acc.:
Profession : 63%
Gender: 87%

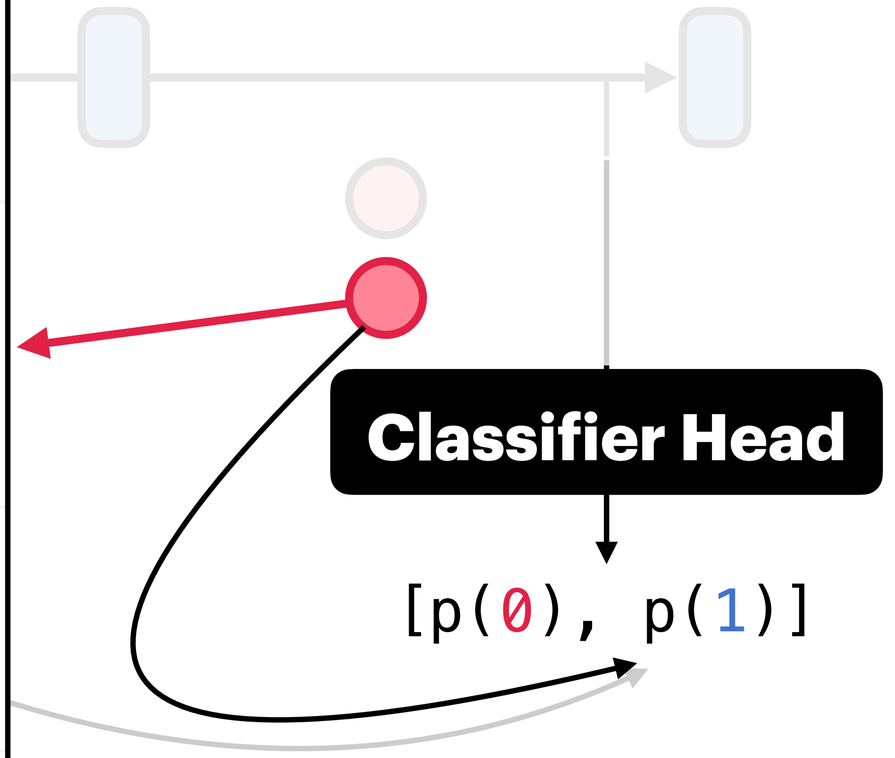
👎

bodies for calf re aring . ↵
↵
It features daily videos of Nicole and Alice ,
along with a few other farmers , doing warm
ups , stretches and strengthening

the marriage was failing . Paul suffered
engulf ing dep ressions . Sometimes he and
Angela barely spoke for days . She felt
swollen with un expressed emotion . “ I

It was like a bitter taste , just a foul taste
,’ he said âĢ¦ Mary Cel este Clement , a
children ’ s book author , lives about 2 miles

At rium at age 13 and that he was preceded in
death by his wife Sarah , who rests next to
him . ↵
↵



Inspect each high-IE feature

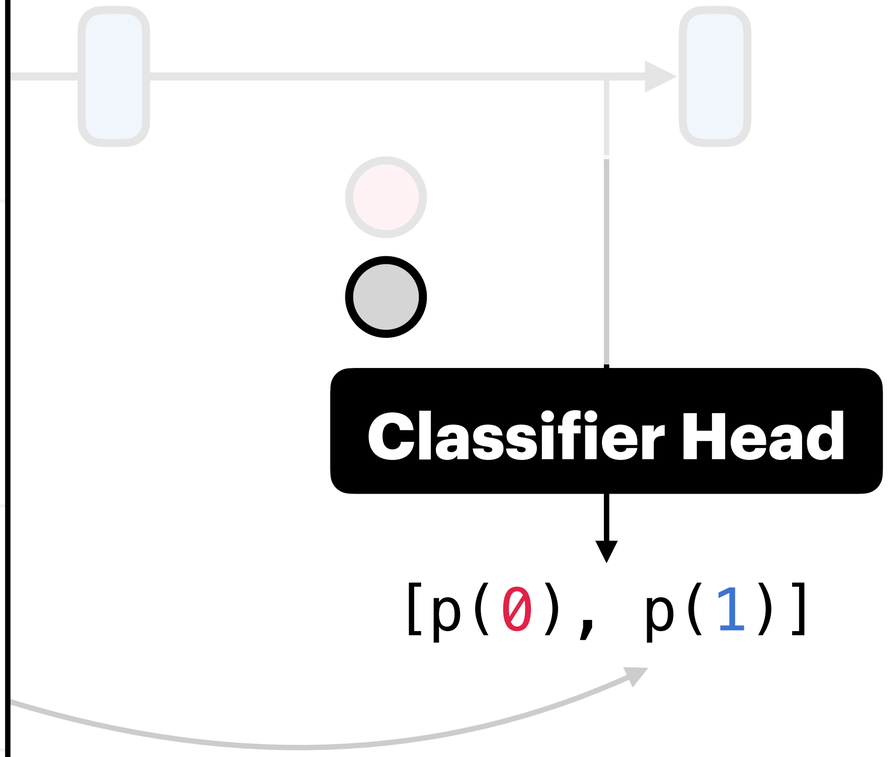
x

bodies for calf re aring . ↵
↵
It features daily videos of Nicole and Alice ,
along with a few other farmers , doing warm
ups , stretches and strengthening

the marriage was failing . Paul suffered
engulf ing dep ressions . Sometimes he and
Angela barely spoke for days . She felt
swollen with un expressed emotion . “ I

It was like a bitter taste , just a foul taste
,’ he said âĢ¦ Mary Cel este Clement , a
children ’ s book author , lives about 2 miles

At rium at age 13 and that he was preceded in
death by his wife Sarah , who rests next to
him . ↵
↵



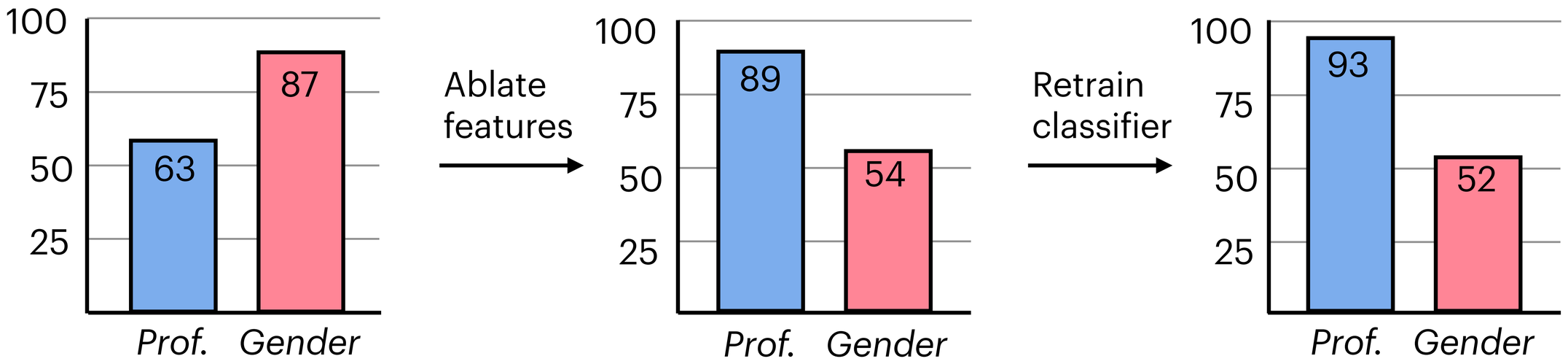
Inspect each high-IE feature

Ablate features that seem related to *gender*

SHIFT

Results

Task: classify profession described in biography



SHIFT

Results

Method	Pythia-70M			Gemma-2-2B		
	↑Profession	↓Gender	↑Worst group	↑Profession	↓Gender	↑Worst group
Original	61.9	87.4	24.4	67.7	81.9	18.2
CBP	83.3	60.1	67.7	90.2	**50.1**	86.7
Random	61.8	87.5	24.4	67.3	82.3	18.0
SHIFT	88.5	54.0	76.0	76.0	51.5	50.0
SHIFT + retrain	**93.1**	**52.0**	**89.0**	**95.0**	52.4	**92.9**

SHIFT

Results

Method	Pythia-70M			Gemma-2-2B		
	↑Profession	↓Gender	↑Worst group	↑Profession	↓Gender	↑Worst group
Original	61.9	87.4	24.4	67.7	81.9	18.2
CBP	83.3	60.1	67.7	90.2	**50.1**	86.7
Random	61.8	87.5	24.4	67.3	82.3	18.0
SHIFT	88.5	54.0	76.0	76.0	51.5	50.0
SHIFT + retrain	**93.1**	**52.0**	**89.0**	**95.0**	52.4	**92.9**
Neuron skyline	75.5	73.2	41.5	65.1	84.3	5.6

Features are a stronger basis than neurons for removing spurious correlations.

SHIFT

Results

Method	Pythia-70M			Gemma-2-2B		
	↑Profession	↓Gender	↑Worst group	↑Profession	↓Gender	↑Worst group
Original	61.9	87.4	24.4	67.7	81.9	18.2
CBP	83.3	60.1	67.7	90.2	**50.1**	86.7
Random	61.8	87.5	24.4	67.3	82.3	18.0
SHIFT	88.5	54.0	76.0	76.0	51.5	50.0
SHIFT + retrain	**93.1**	**52.0**	**89.0**	**95.0**	52.4	**92.9**
Neuron skyline	75.5	73.2	41.5	65.1	84.3	5.6
Feature skyline	88.5	54.3	62.9	80.8	53.7	56.7

Features are a stronger basis than neurons for removing spurious correlations.

Our judgments about feature relevance are largely informative.

SHIFT

Results

Method	Pythia-70M			Gemma-2-2B		
	↑Profession	↓Gender	↑Worst group	↑Profession	↓Gender	↑Worst group
Original	61.9	87.4	24.4	67.7	81.9	18.2
CBP	83.3	60.1	67.7	90.2	**50.1**	86.7
Random	61.8	87.5	24.4	67.3	82.3	18.0
SHIFT	88.5	54.0	76.0	76.0	51.5	50.0
SHIFT + retrain	**93.1**	**52.0**	**89.0**	**95.0**	52.4	**92.9**
Neuron skyline	75.5	73.2	41.5	65.1	84.3	5.6
Feature skyline	88.5	54.3	62.9	80.8	53.7	56.7
Oracle	93.0	49.4	91.9	95.0	50.6	93.1

Features are a stronger basis than neurons for removing spurious correlations.

Our judgments about feature relevance are largely informative.

*SHIFT achieve the performance of a classifier trained on **unbiased** data!*

Findings using counterfactual interventions

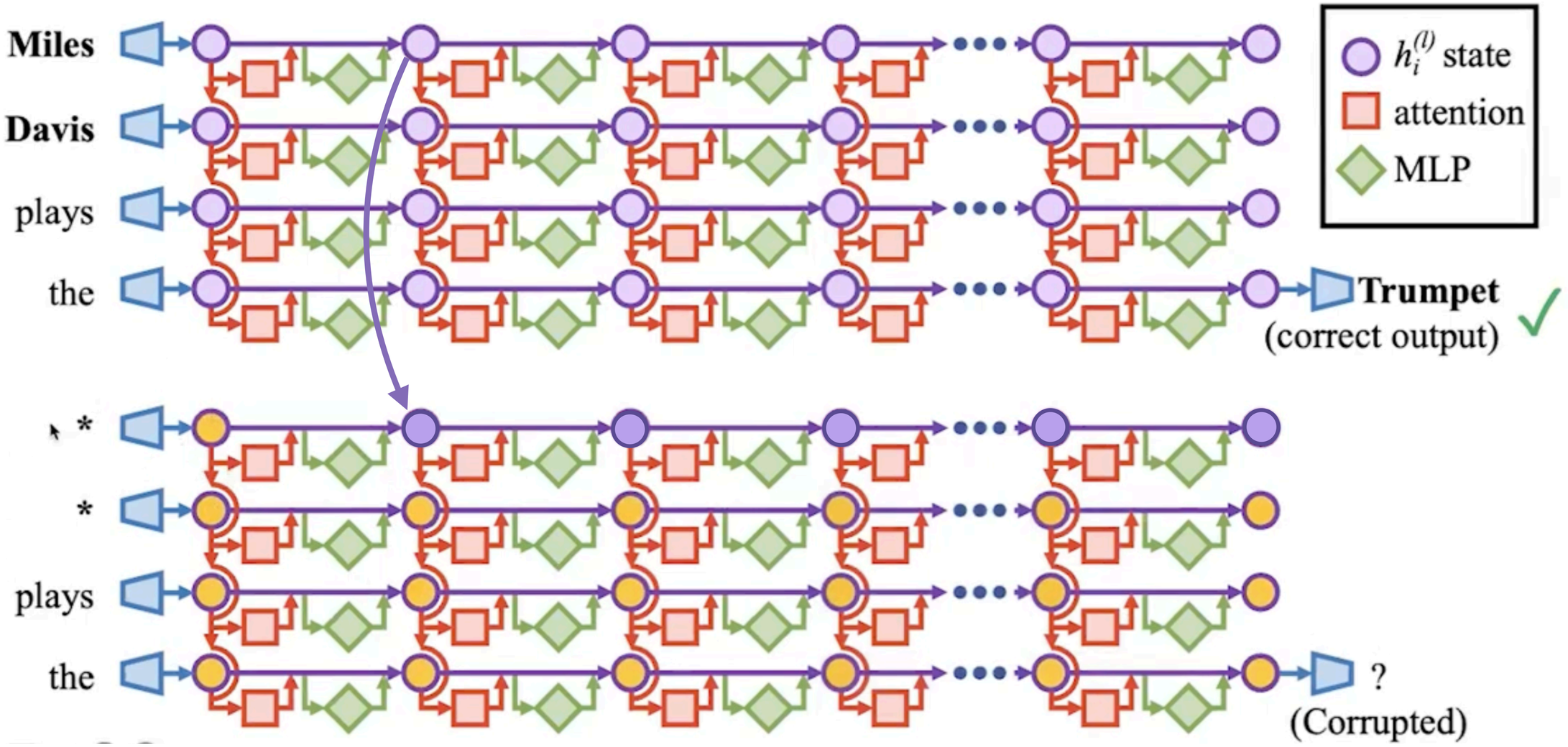
- Language models—even very small ones—implement human-like mechanisms for resolving syntactic dependencies **[Marks et al., 2024]**
- Models neither memorize nor implement robust algorithms to perform arithmetic; they implement "bags of heuristics" **[Nikankin et al., 2024]**
- Models can be precisely edited to remove gender bias and *improve performance* **[Marks et al., 2024]** and to change factual knowledge **[Meng et al., 2022; 2023]**
- Models can be steered to use more or less angry language, to disregard errors in code, to talk more about certain topics, to adopt different styles, etc., without significantly harming output quality **[Templeton et al., 2024]**
- Models can be encouraged to learn particular mechanisms or abstractions by design **[Geiger et al., 2022]**

Updating Knowledge in LMs

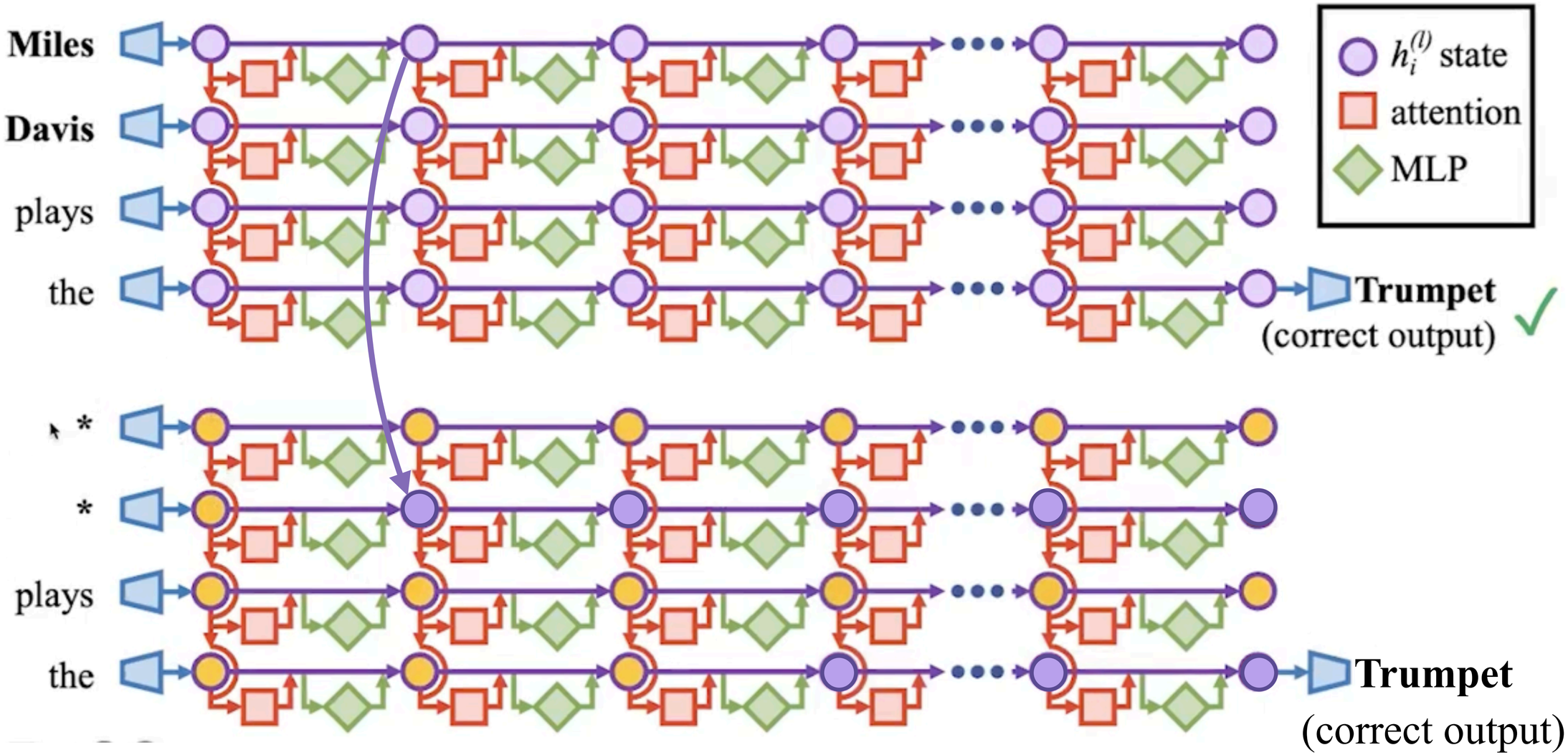
- Now, instead of removing info from the model, we'd like to *add* or *edit* knowledge in the LM's parameters
- This will require a very different approach: sparse features and classifiers can't add information that wasn't already in the model to some extent
- Fine-tuning on updated info might work, but it's coarse and could have many unintended side effects

- Let's do targeted model editing!

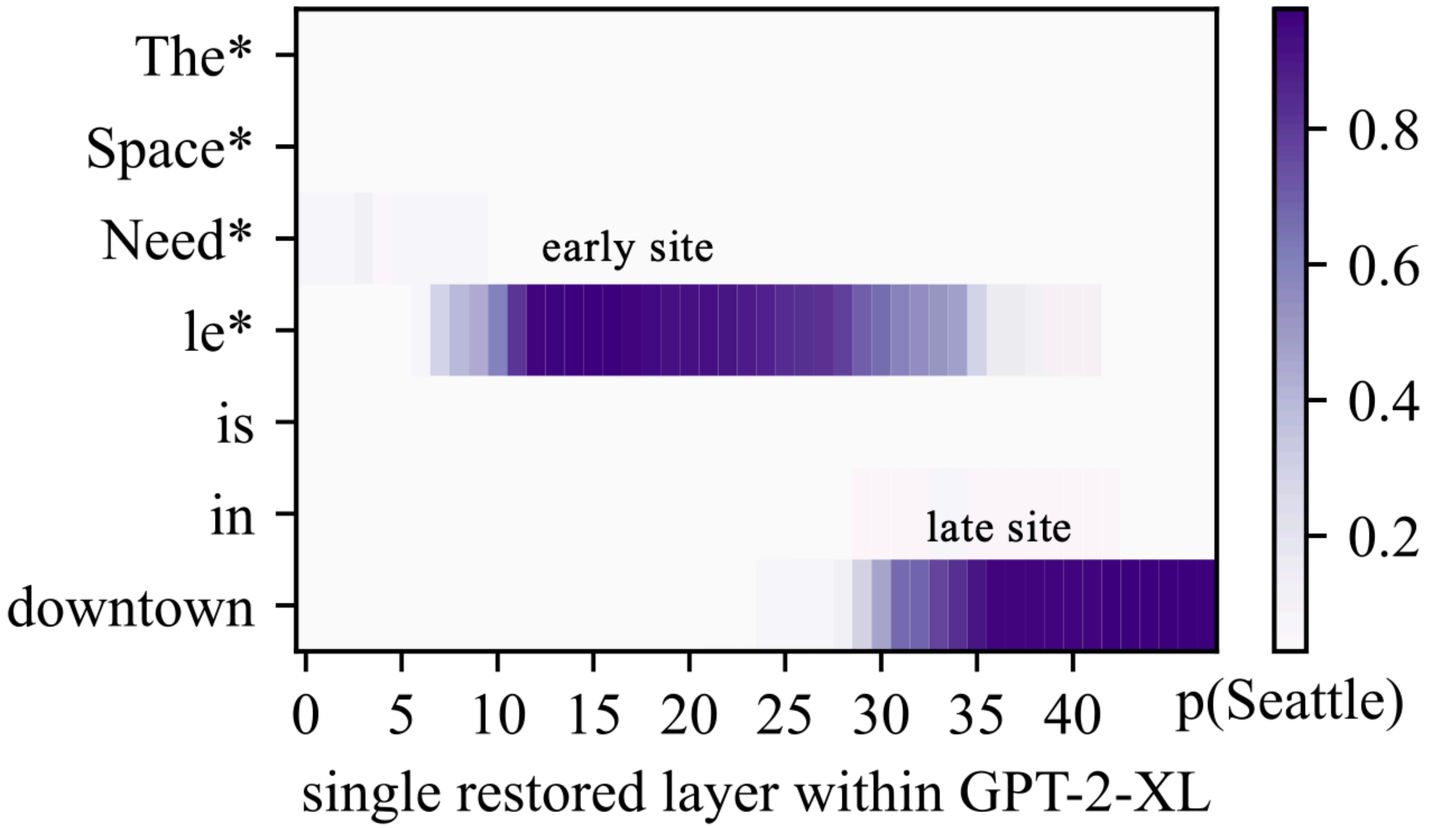
Locating Knowledge in LMs



Locating Knowledge in LMs



Causal Tracing



If we do this across all layers and token positions, we have two main sites of causal importance:

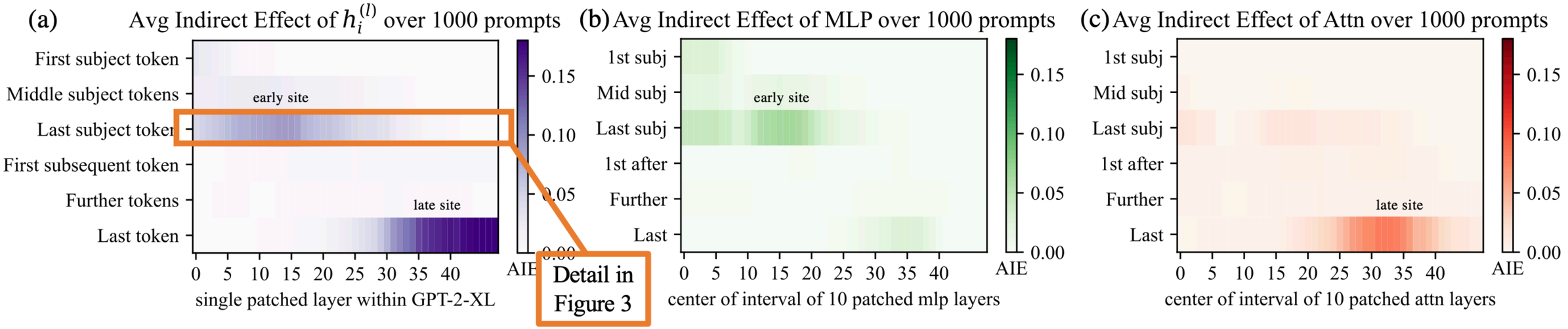
The **early site**

The **late site**

What's happening in these sites?

Hypothesis: early site is knowing, late site is just saying

Causal Tracing



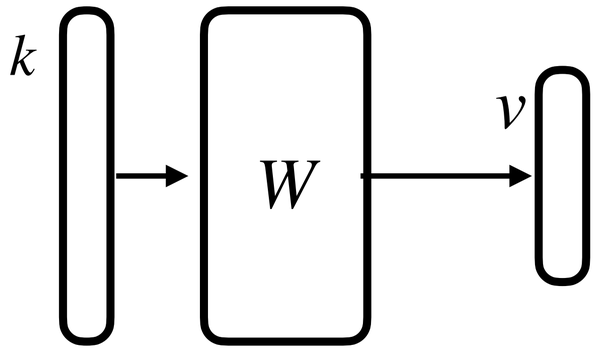
- Knowledge is encoded quite consistently across facts:
  - Two sites
  - Last subject token
  - Mid-layer MLP and late-layer attention modules do a lot of the work

Associative Memory View of a Neural Layer

- Why are we able to find knowledge like this?
- You can think of a layer as a key-value memory:

$$\{k_1 \rightarrow v_1, k_2 \rightarrow v_2, \ldots, k_n \rightarrow v_n\}$$

- Where k_i is the subject, and v_i is the completion of the fact.



You can think of an MLP as learning weights W s.t. $Wk_i \approx v_i$

Editing Knowledge in a Neural Layer

- Goal: replace a fact with a new one $k_* \rightarrow v_*$ while minimizing error on the old facts $K_0 \rightarrow V_0$.

$$V_1 = \begin{bmatrix} V_0 & v_* \end{bmatrix} \qquad K_1 = \begin{bmatrix} K_0 & k_* \end{bmatrix}$$

- Basically, we want to change the model weights W_0 to W_1 s.t. they encode the new fact while minimally modifying the old ones.

$$W_1 = \arg\min_W \| V_1 - W K_1 \|^2$$

- This is a **least-squares** problem, which has a closed-form solution:

$$W_1 K_0 K_0^\top + W_1 k_* k_*^\top = V_0 K_0^\top + v_* k_*^\top$$

Editing Knowledge in a Neural Layer

$$W_1 K_0 K_0^\top + W_1 k_* k_*^\top = V_0 K_0^\top + v_* k_*^\top$$

- Let's subtract the original least-squares solution from the new solution:

$$W_0 K_0 K_0^\top = V_0 K_0^\top$$

output according to old weights

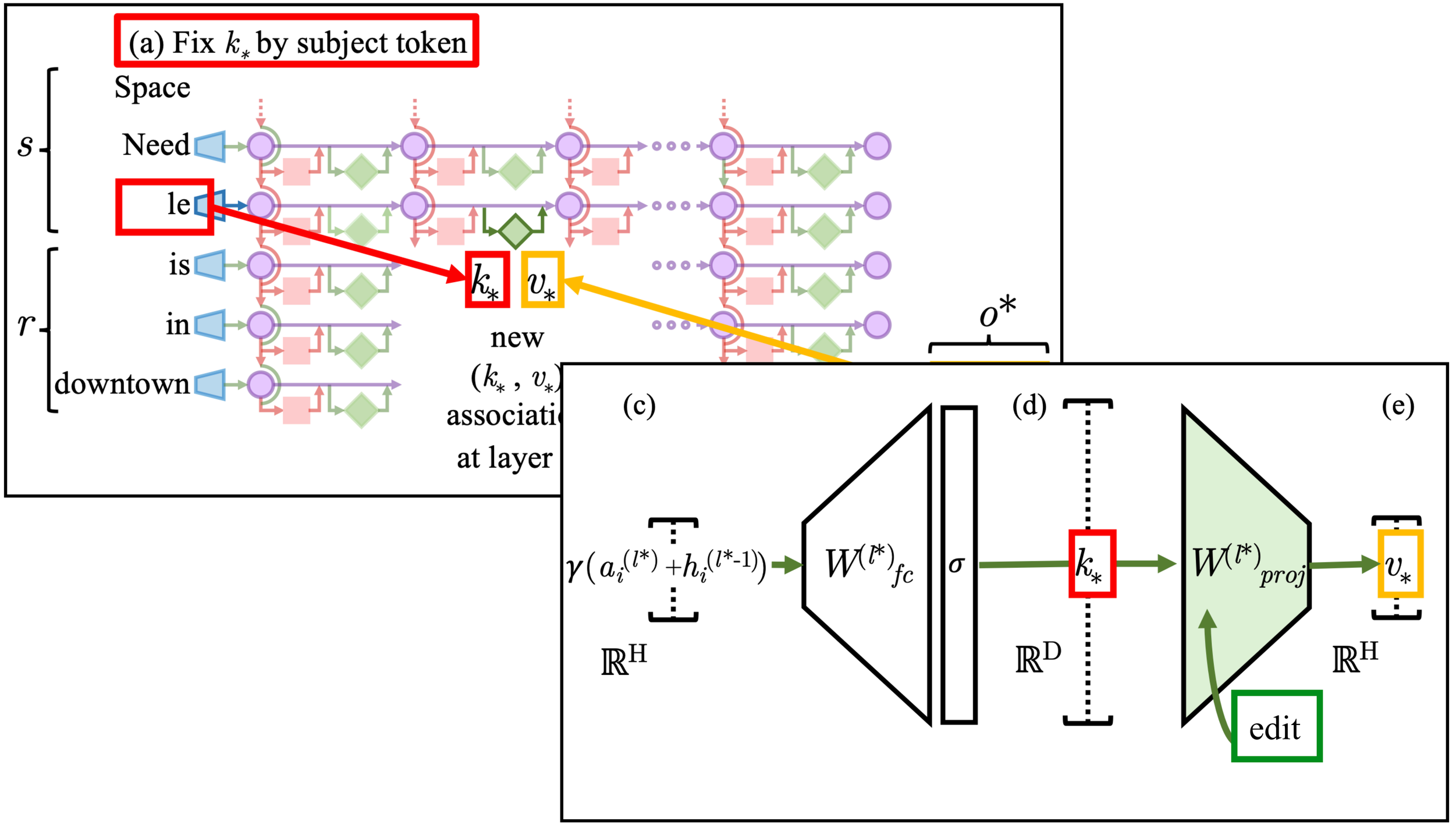
desired output

- Define: $r = v_* - W_0 k_*^\top, \quad C_0 = K_0 K_0^\top$ covariance of original keys

new weights $W_1 = W_0 + r k_*^\top (C_0 + k_* k_*^\top)^{-1}$

old weights

covariance of new key


(a) Fix k_* by subject token
Space
Need
le
is
in
downtown
s
r
k_*
v_*
o^*
new
(k_*, v_*)
associati
at layer
(c)
(d)
(e)
$\gamma(a_i^{(l^*)} + h_i^{(l^*-1)})$
$W^{(l^*)}_{fc}$
σ
k_*
$W^{(l^*)}_{proj}$
v_*
$\mathbb{R}^H$
$\mathbb{R}^D$
$\mathbb{R}^H$
edit

Evaluation

- *How do we evaluate the success of our edits?*
  - **Generalization:** knowledge is consistent under rephrases
  - **Specificity:** different types of knowledge don’t interfere with each other

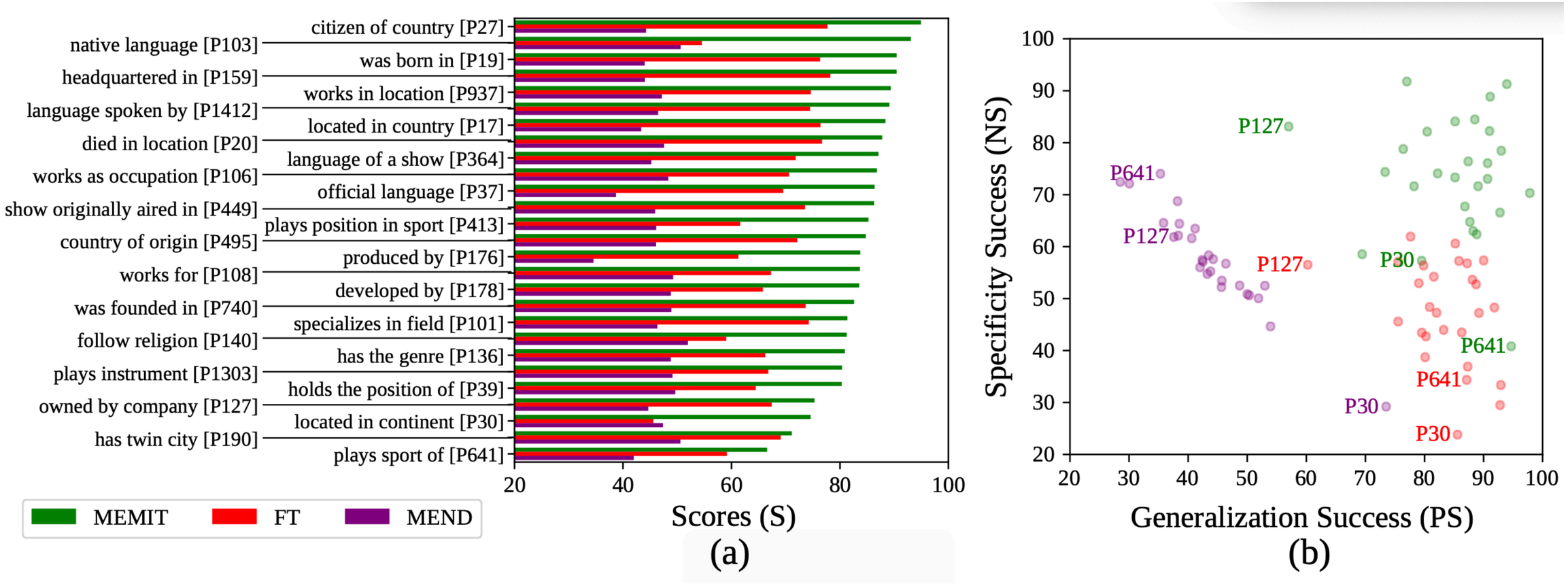
<u>Added fact: The Eiffel Tower is in Rome.</u>

The city where the Eiffel tower is located is... (paraphrase generalization)

How can I get to the Eiffel Tower? (consistency generalization)

What is there to eat near the Eiffel Tower? (consistency generalization)

Where is the Sears Tower? (specificity)



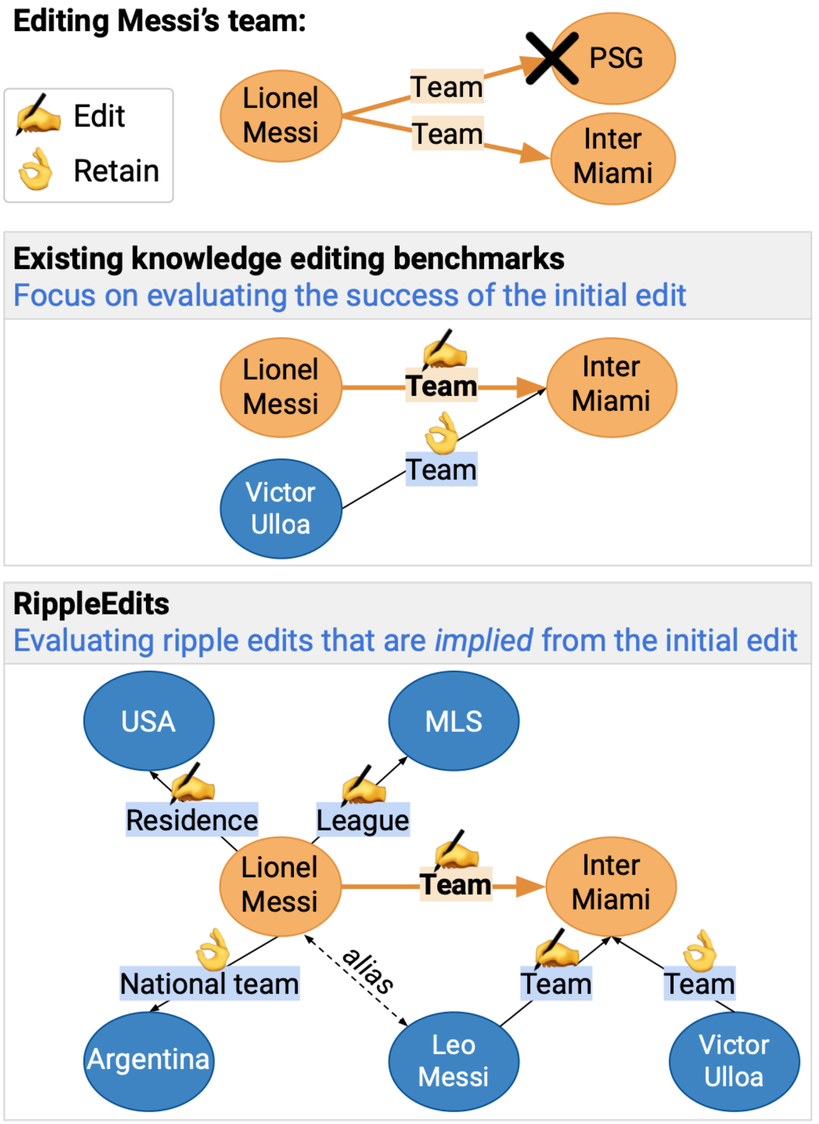
Knowledge editing (MEMIT) seems to work better than fine-tuning (FT)!

There is a trade-off between generalization and specificity.

Known Challenges in Model Editing

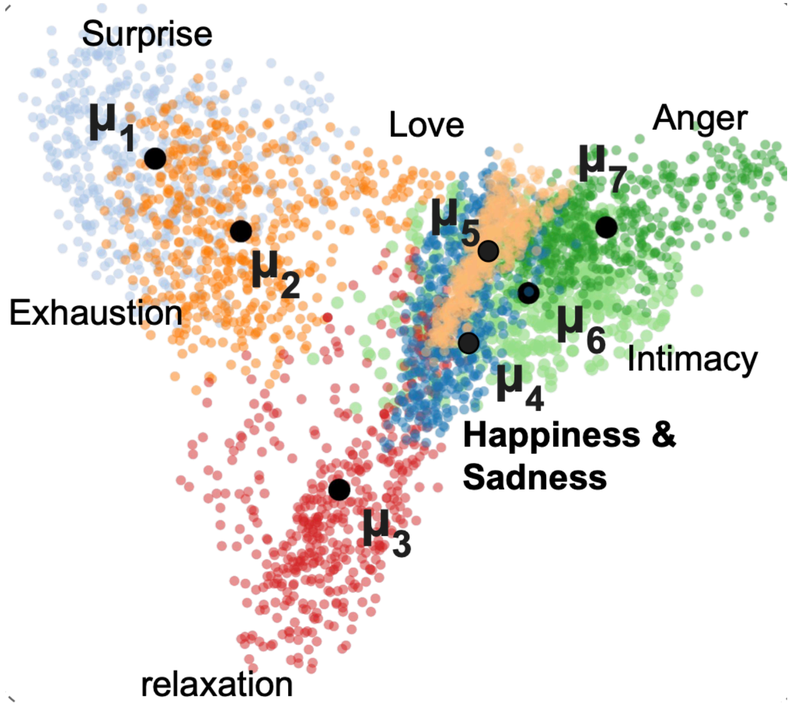
- It's been found that edits often don't generalize to logically related facts.

- Many folks now conceptualize model editing as operating over bigram probabilities, rather than latent knowledge *per se*.
    - (This is a bit pessimistic, but many failure modes could be explained by this perspective.)



The Near Future of Interpretability

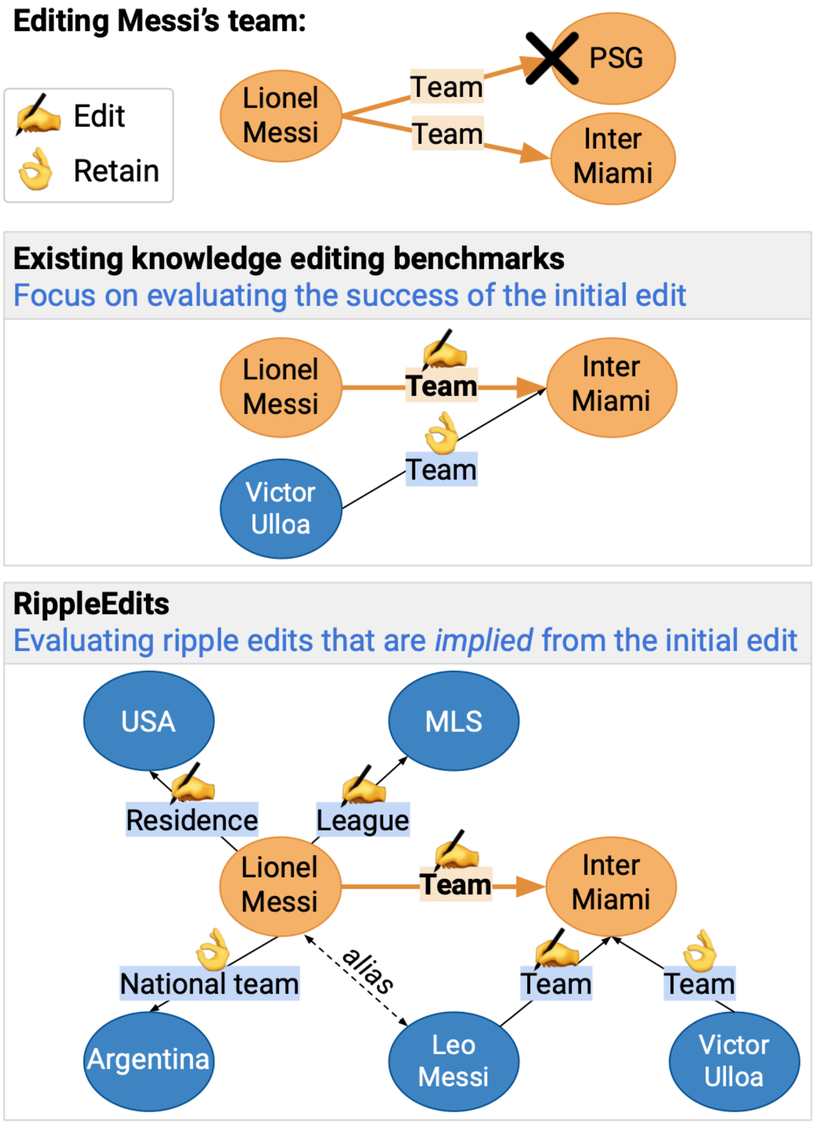
- Discovering better abstractions/featurizers
  - Multi-dimensional features? Non-binary features?



The Near Future of Interpretability

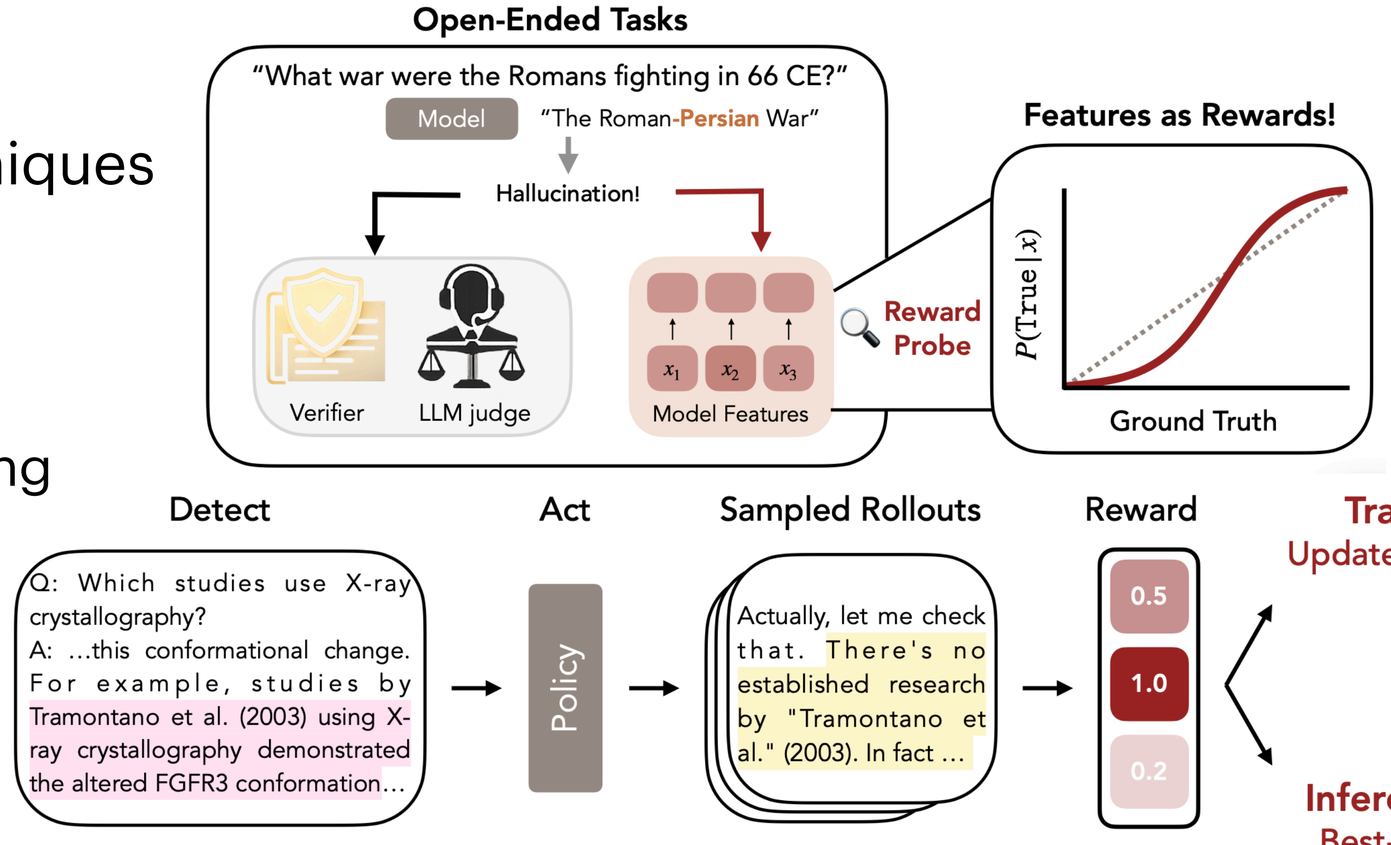
- Discovering better abstractions/featurizers
  - Multi-dimensional features? Non-binary features?

- More effective model editing techniques



The Near Future of Interpretability

- Discovering better abstractions/featurizers
  - Multi-dimensional features? Non-binary features?
- More effective model editing techniques
- Pragmatic interpretability
  - Fixing models during (post-)training
  - Monitoring activations



More References and Resources

- Surveys on interpretability methods:
  - The Quest for the Right Mediator: Surveying Mechanistic Interpretability for NLP through the Lens of Causal Mediation Analysis **[Mueller et al., 2026]**
  - Probing Classifiers: Promises, Shortcomings, and Advances **[Belinkov, 2021]**
  - Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability **[Geiger et al., 2025]**

- Play around with and try to interpret sparse features: Neuronpedia **[Lin & Bloom, 2023]**
- ROME **[Meng*, Bau* et al., 2022]** and MEMIT **[Meng et al., 2023]**

Next Week

- Retrieval and tool use
  - Retrieval-augmented generation (RAG) systems
  - Environment understanding
  - Agentic workflows
  - Prompt induction and tool induction

- AI safety, bias, and ethics