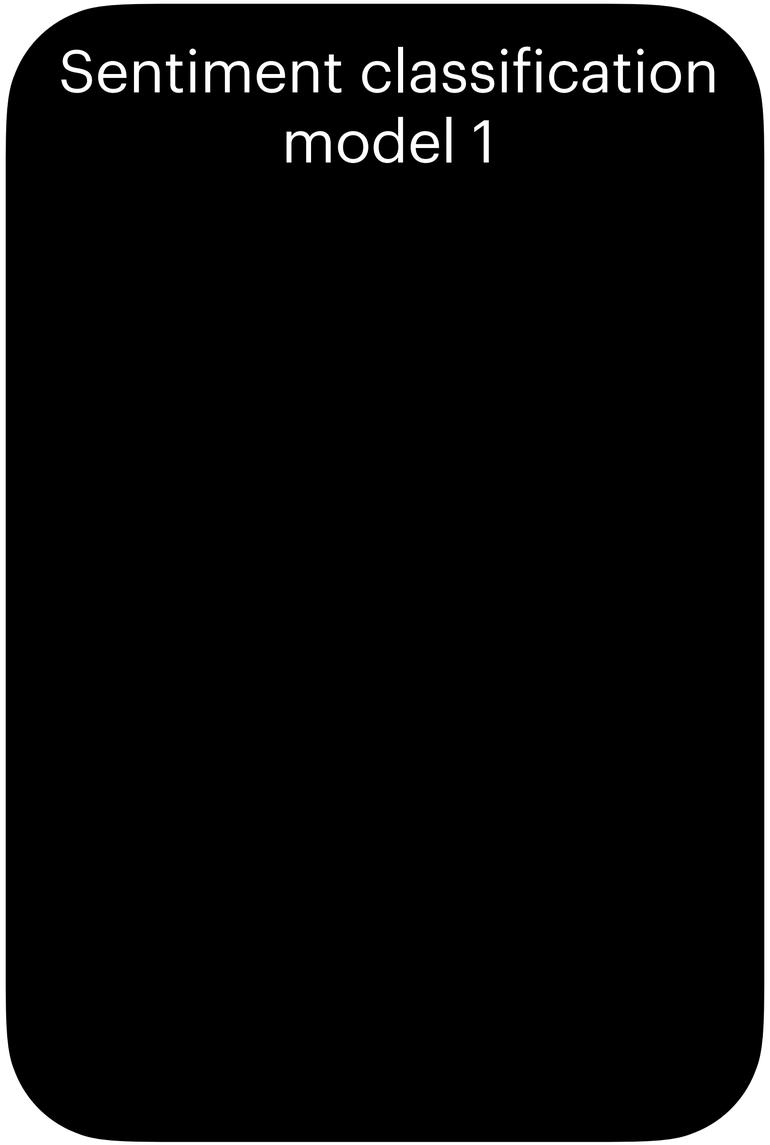
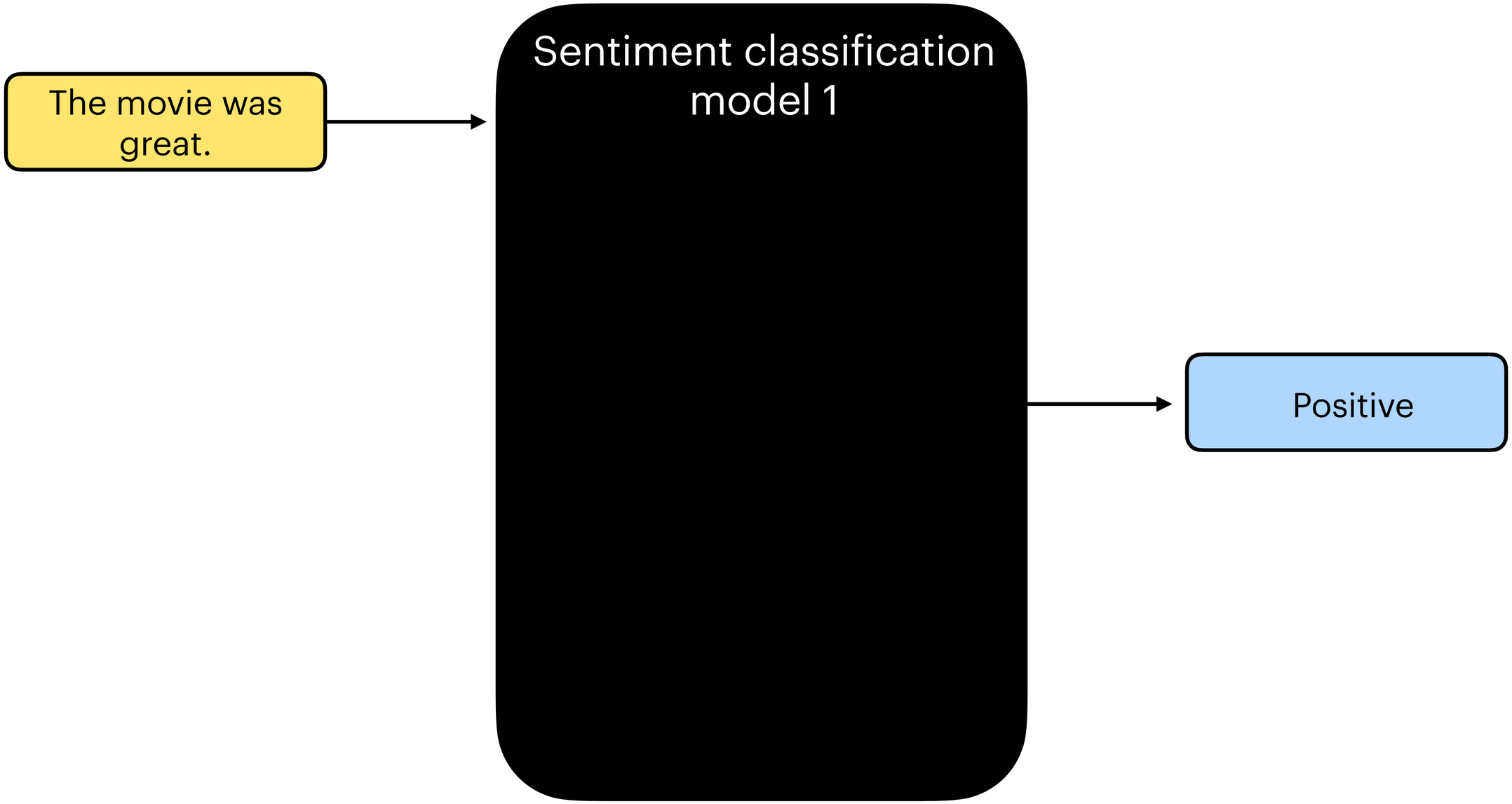
Interpretability I

Aaron Mueller
CS599 B1: Advanced Natural Language Processing
Nov. 5, 2026

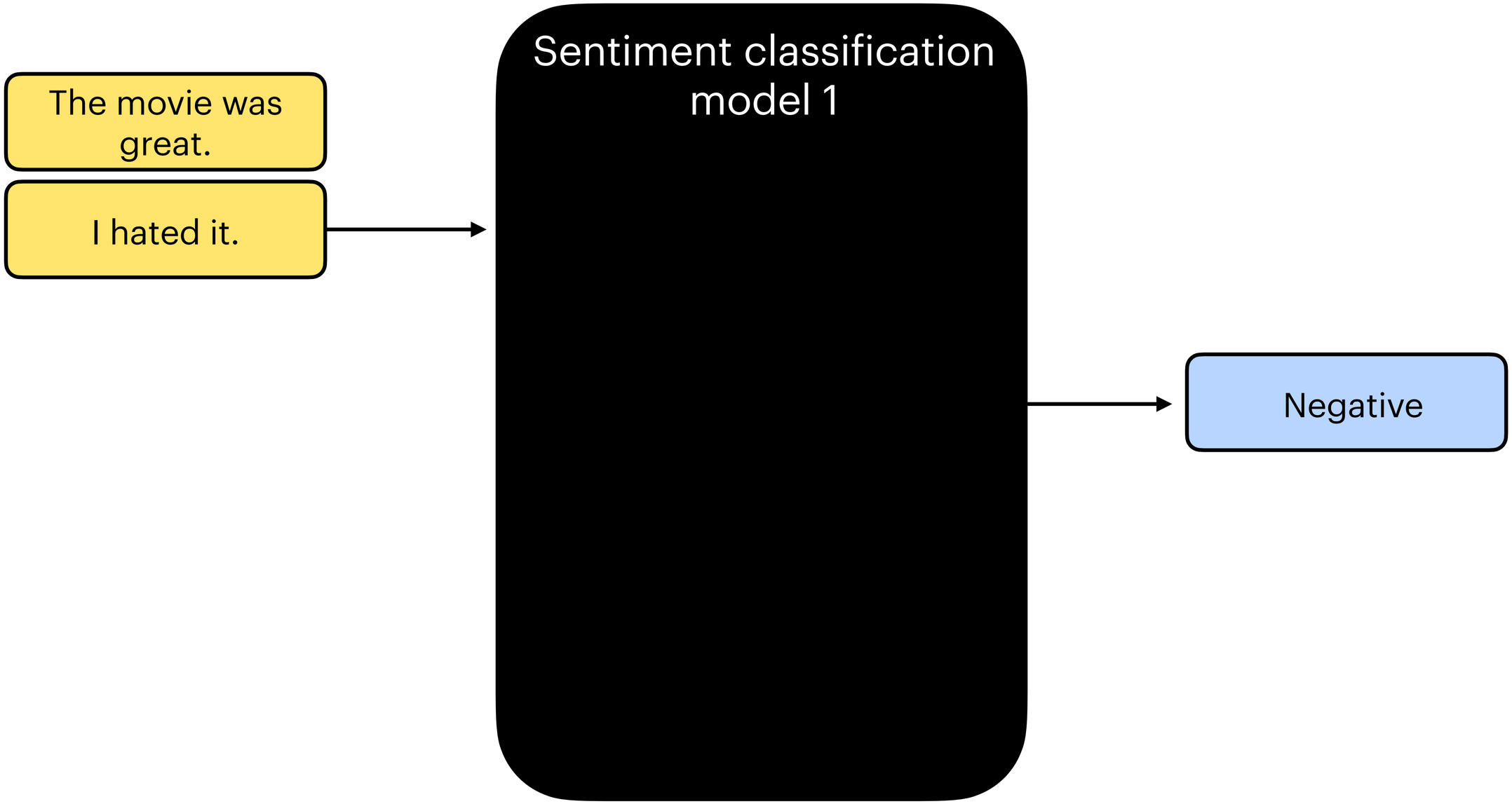
The Limitations of Behavioral Testing



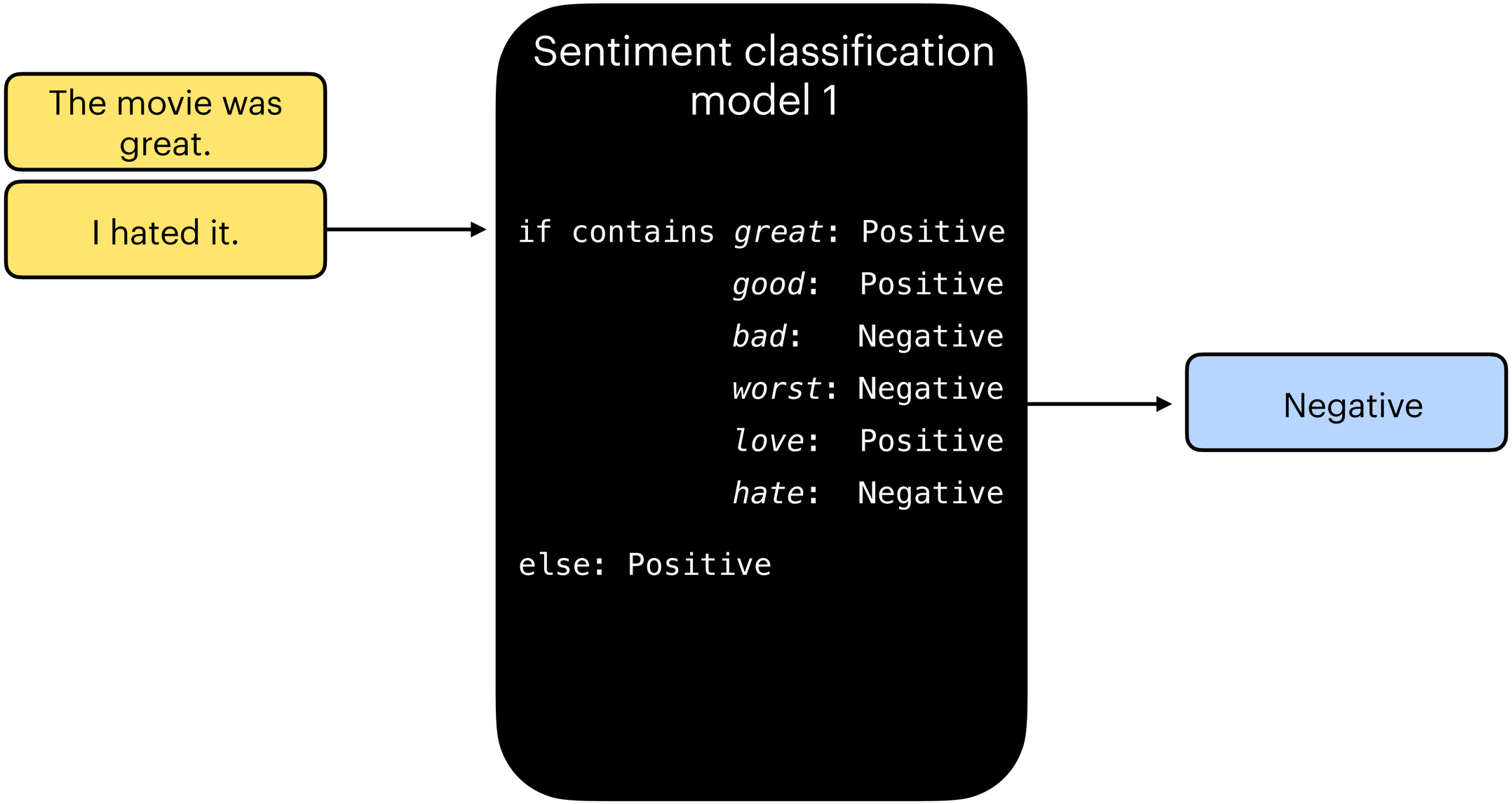
The Limitations of Behavioral Testing



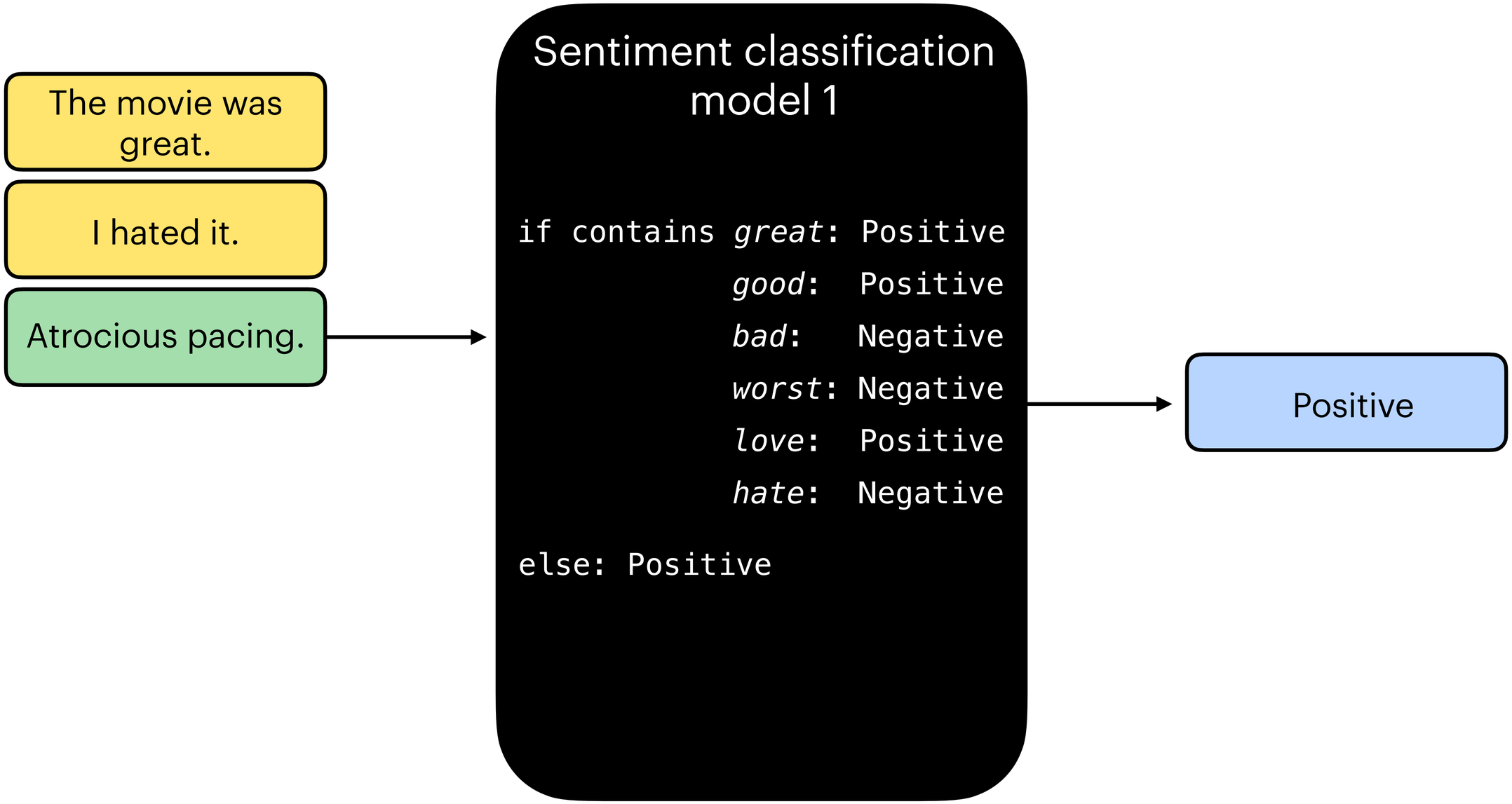
The Limitations of Behavioral Testing



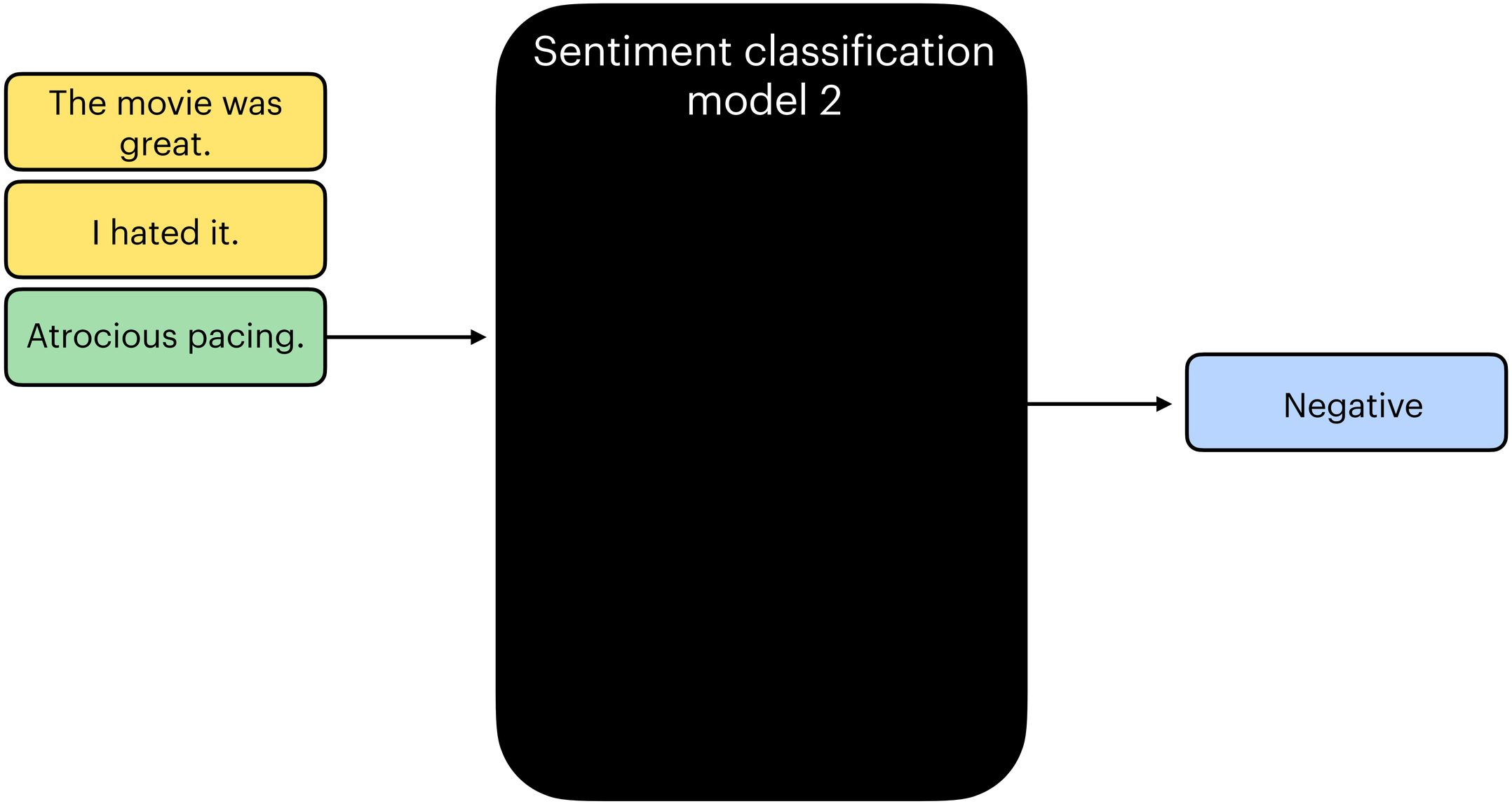
The Limitations of Behavioral Testing



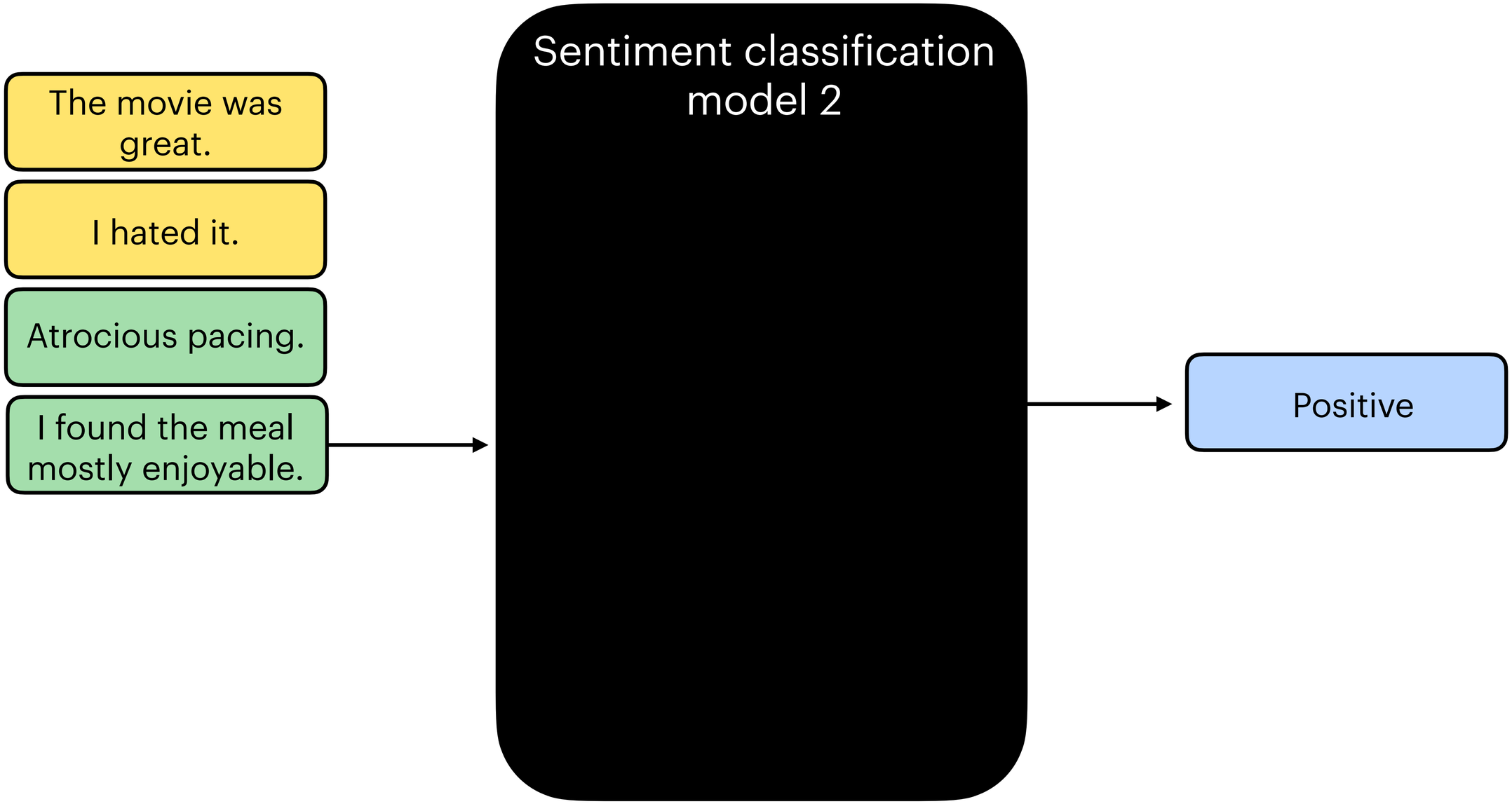
The Limitations of Behavioral Testing



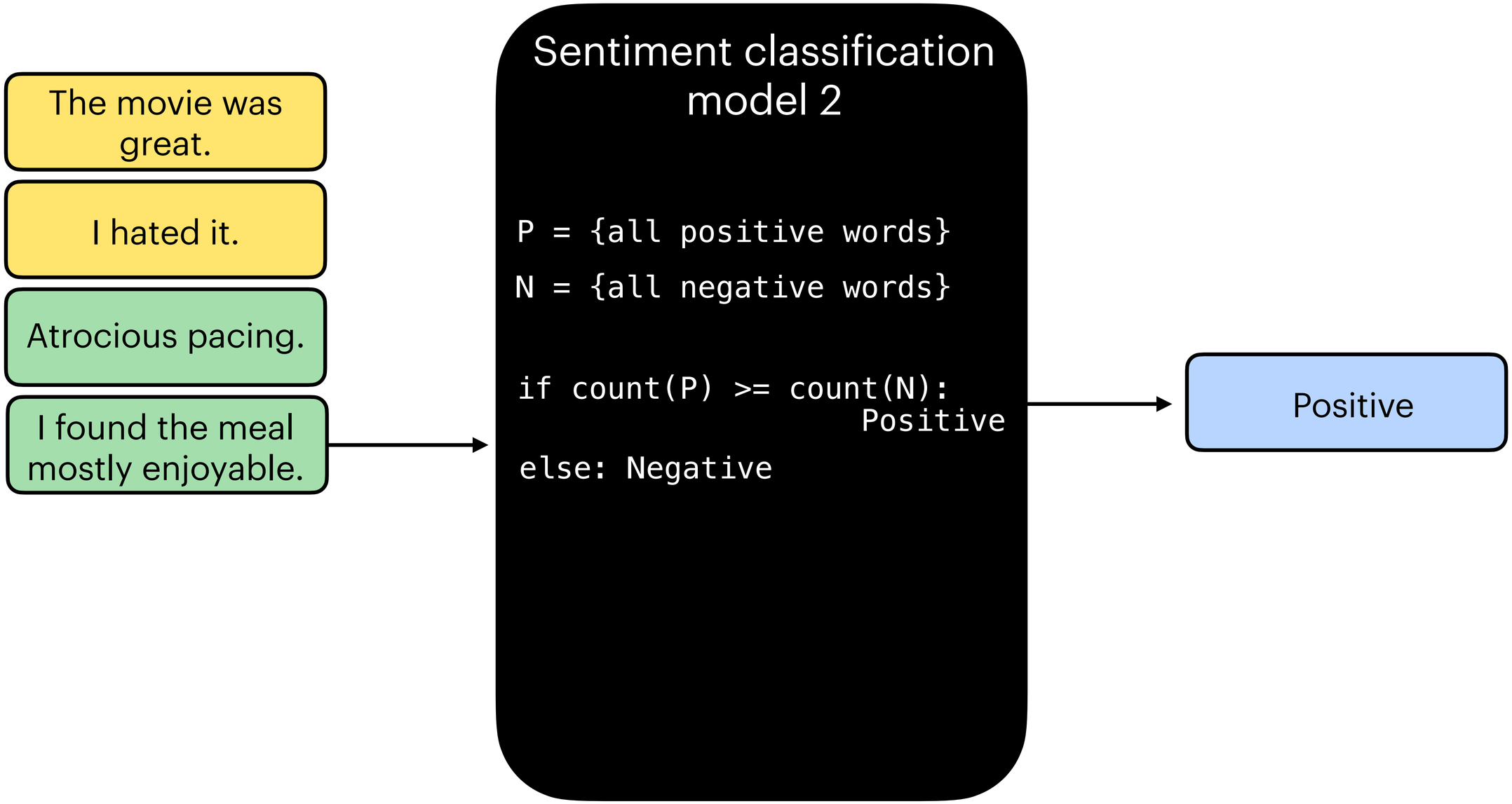
The Limitations of Behavioral Testing



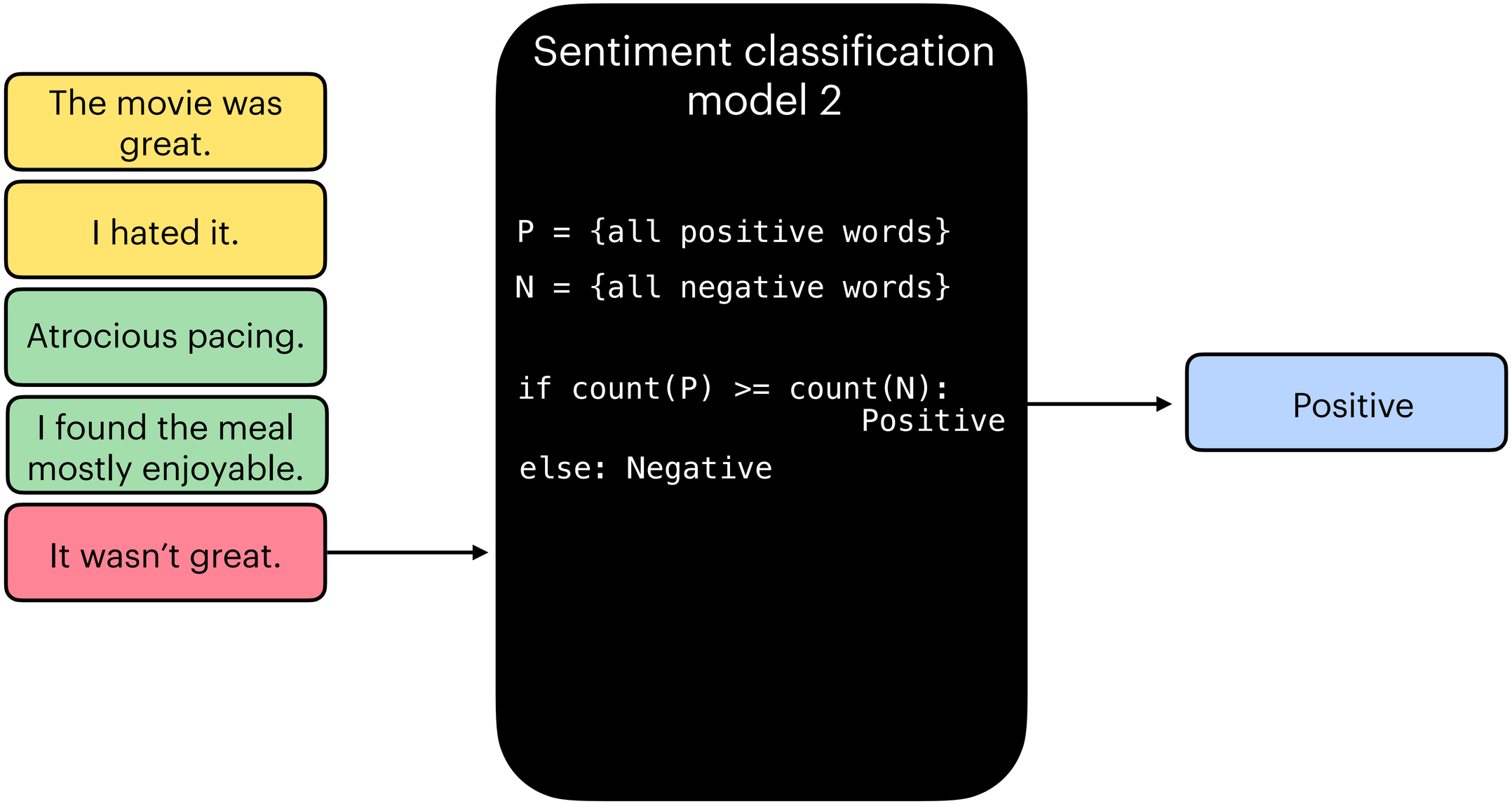
The Limitations of Behavioral Testing



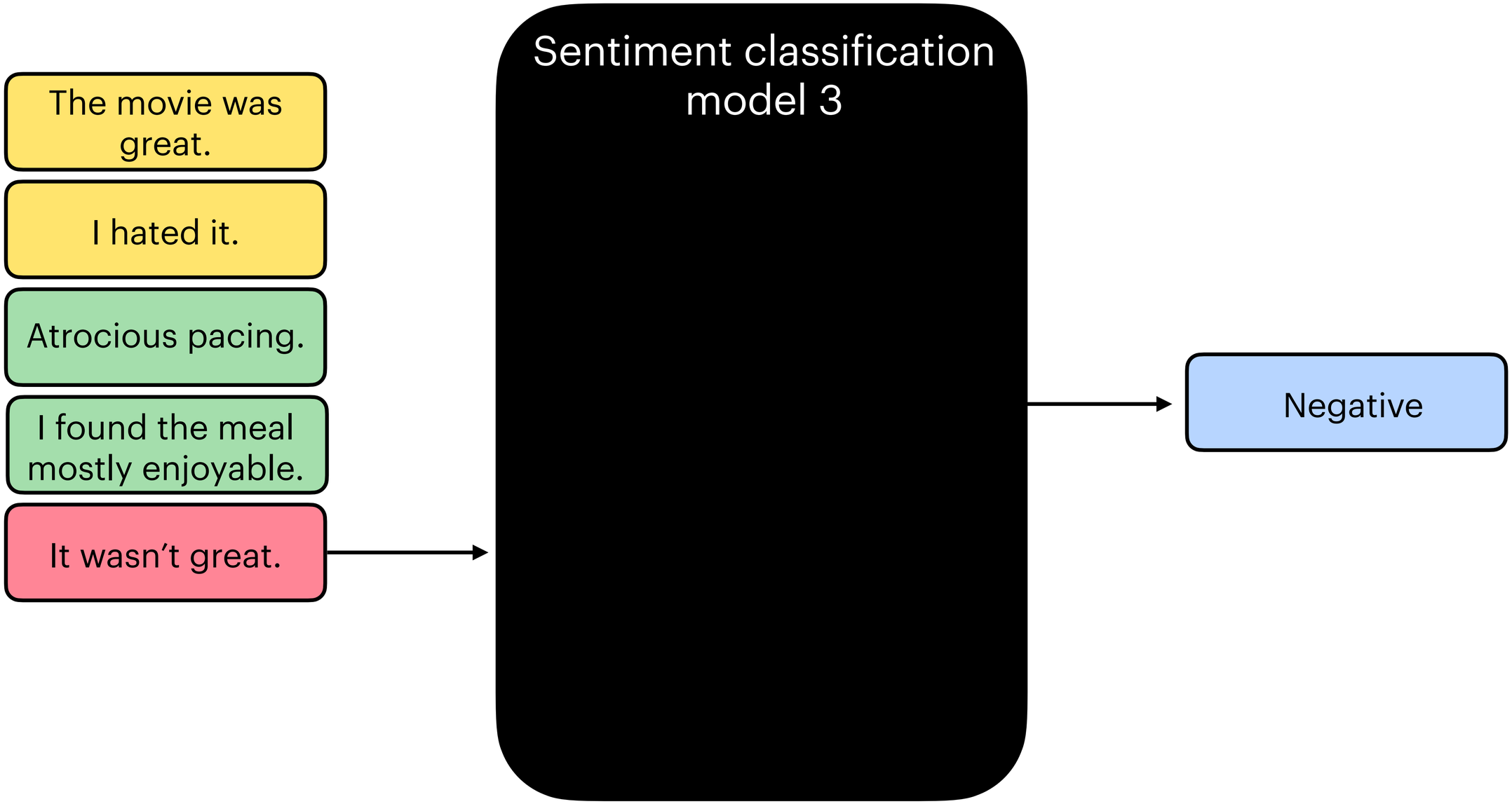
The Limitations of Behavioral Testing



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?

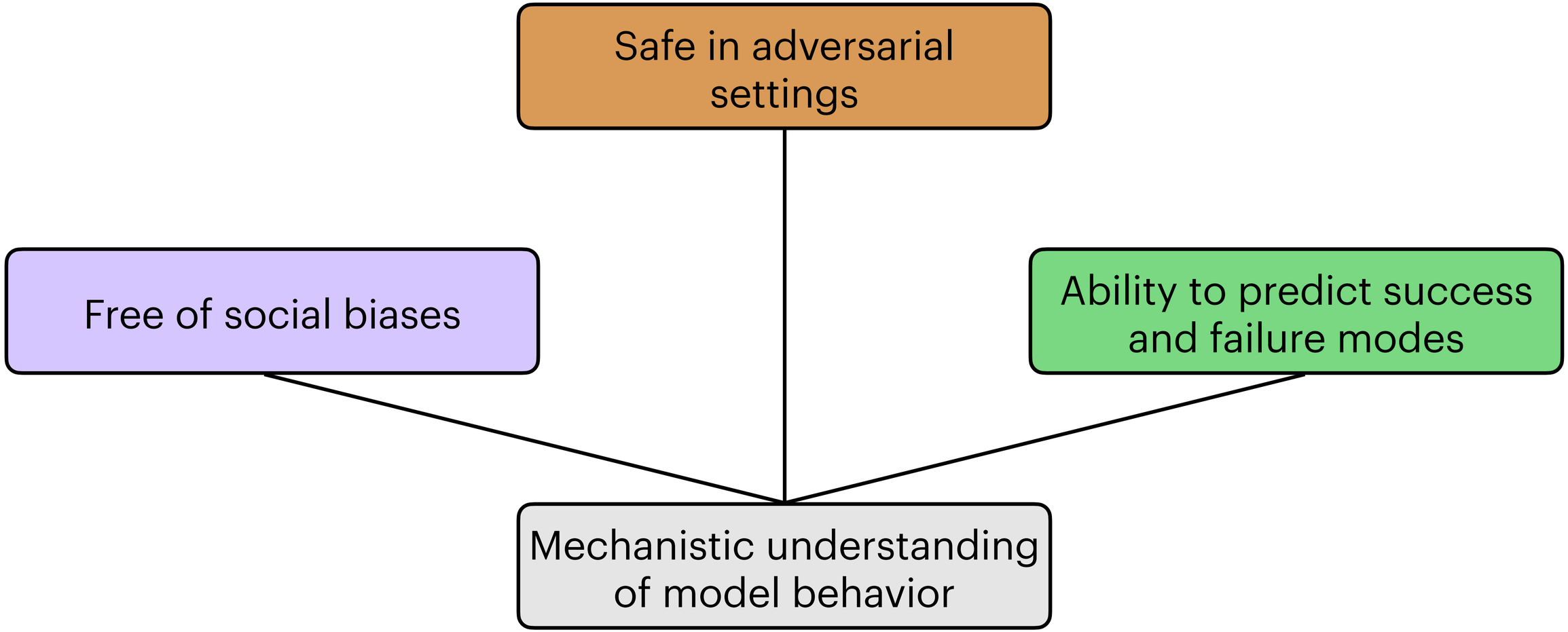
Critical Goals

Safe in adversarial settings

Free of social biases

Ability to predict success and failure modes

Critical Goals



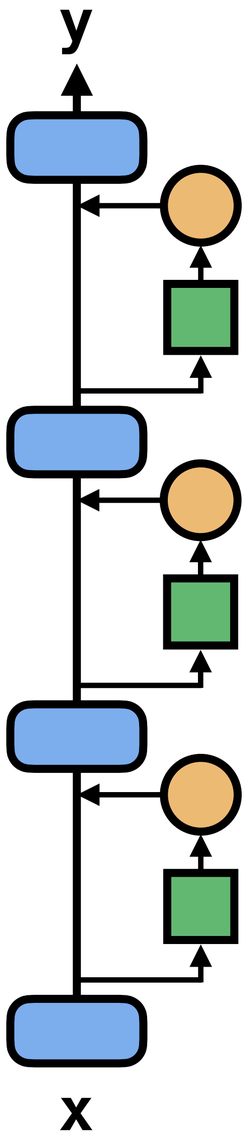
Outline

1. Probing

2. Localization: causal mediation analysis

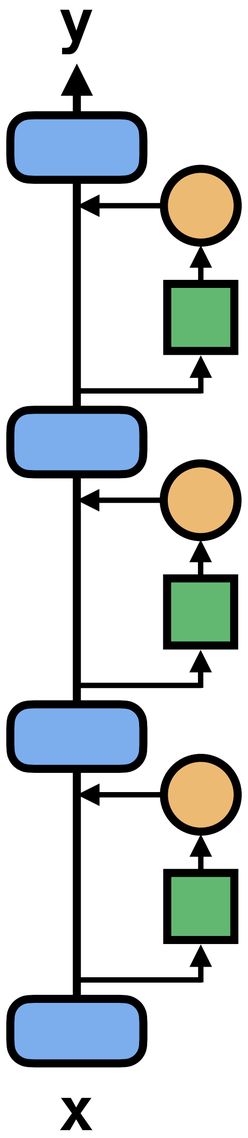
3. Application: model editing

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

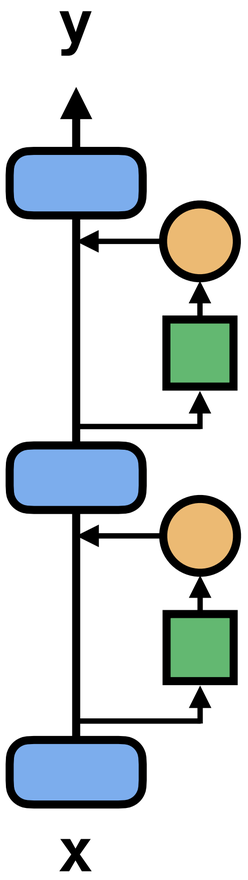
1. Hypothesize some feature, concept, or internal structure encoded by the model.

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.

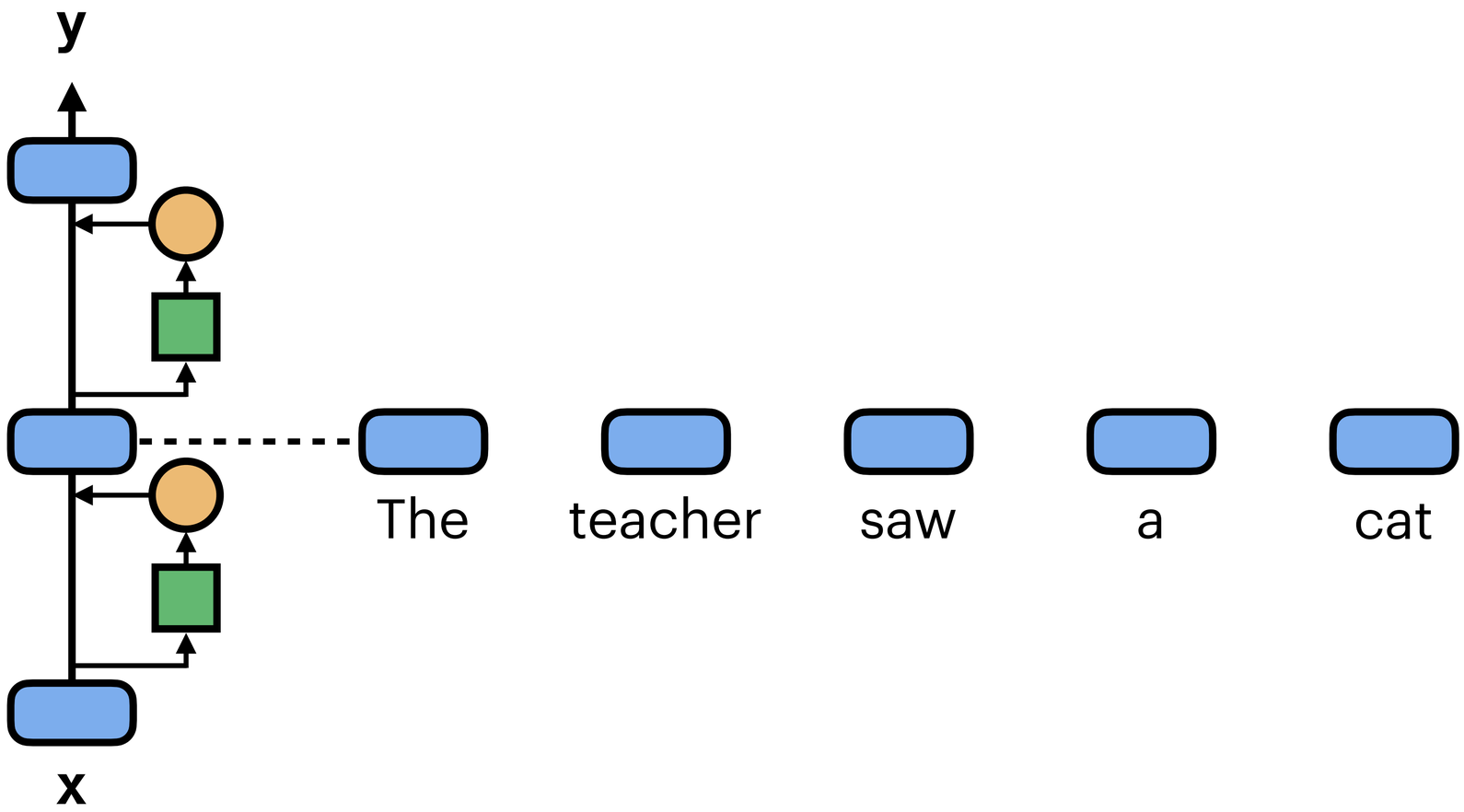
Probing





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

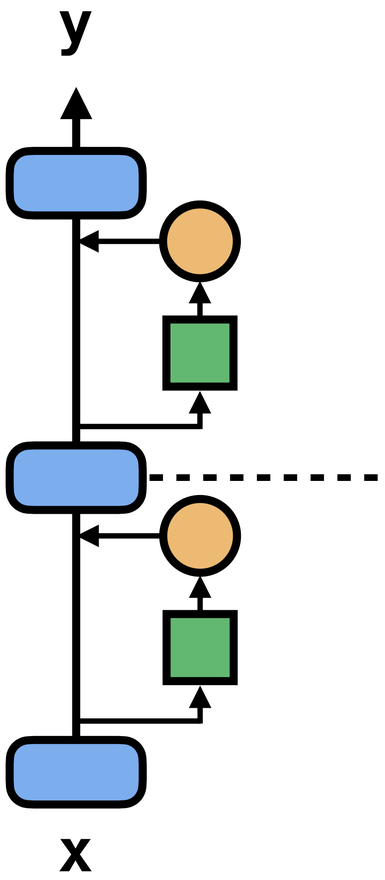
Probing



attention MLP
residual vector

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

Probing



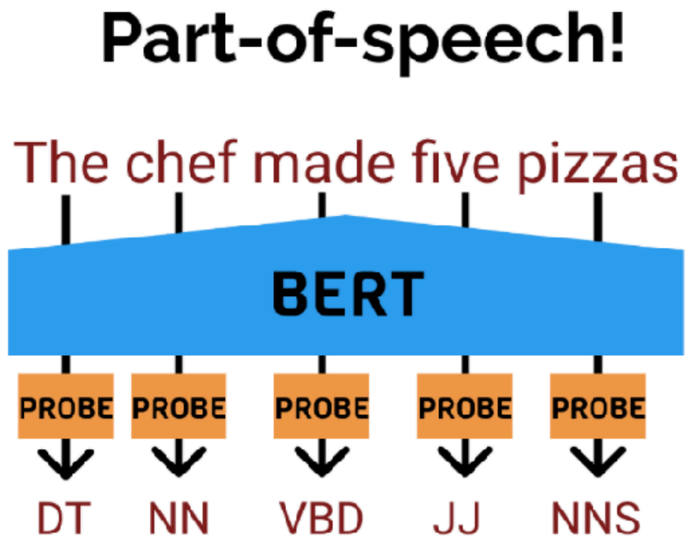
probe = SmallLinearClassifier(X, y)

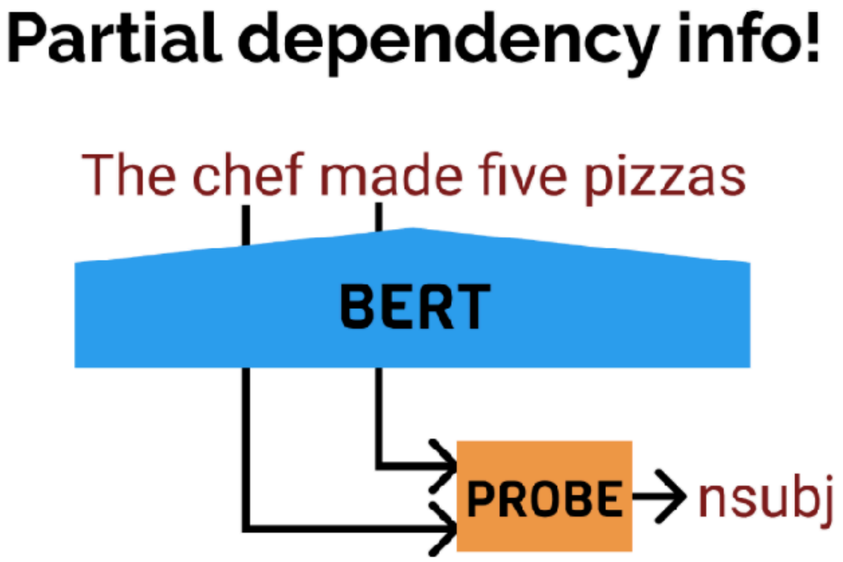
h_1	h_2	h_3	h_4	h_5
The	teacher	saw	a	cat
DET	NOUN	VERB	DET	NOUN

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

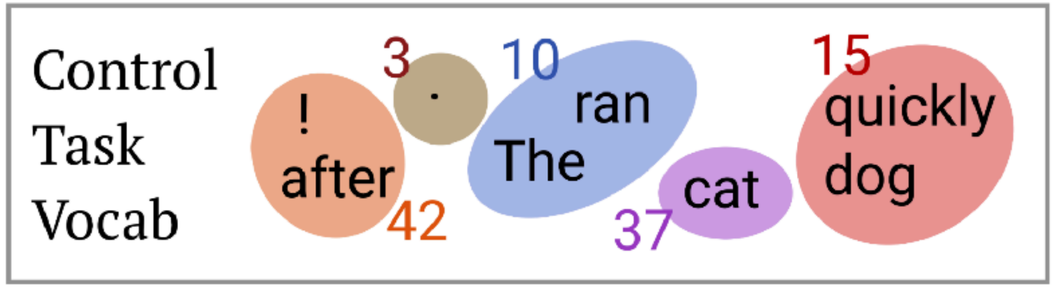
Control Task

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

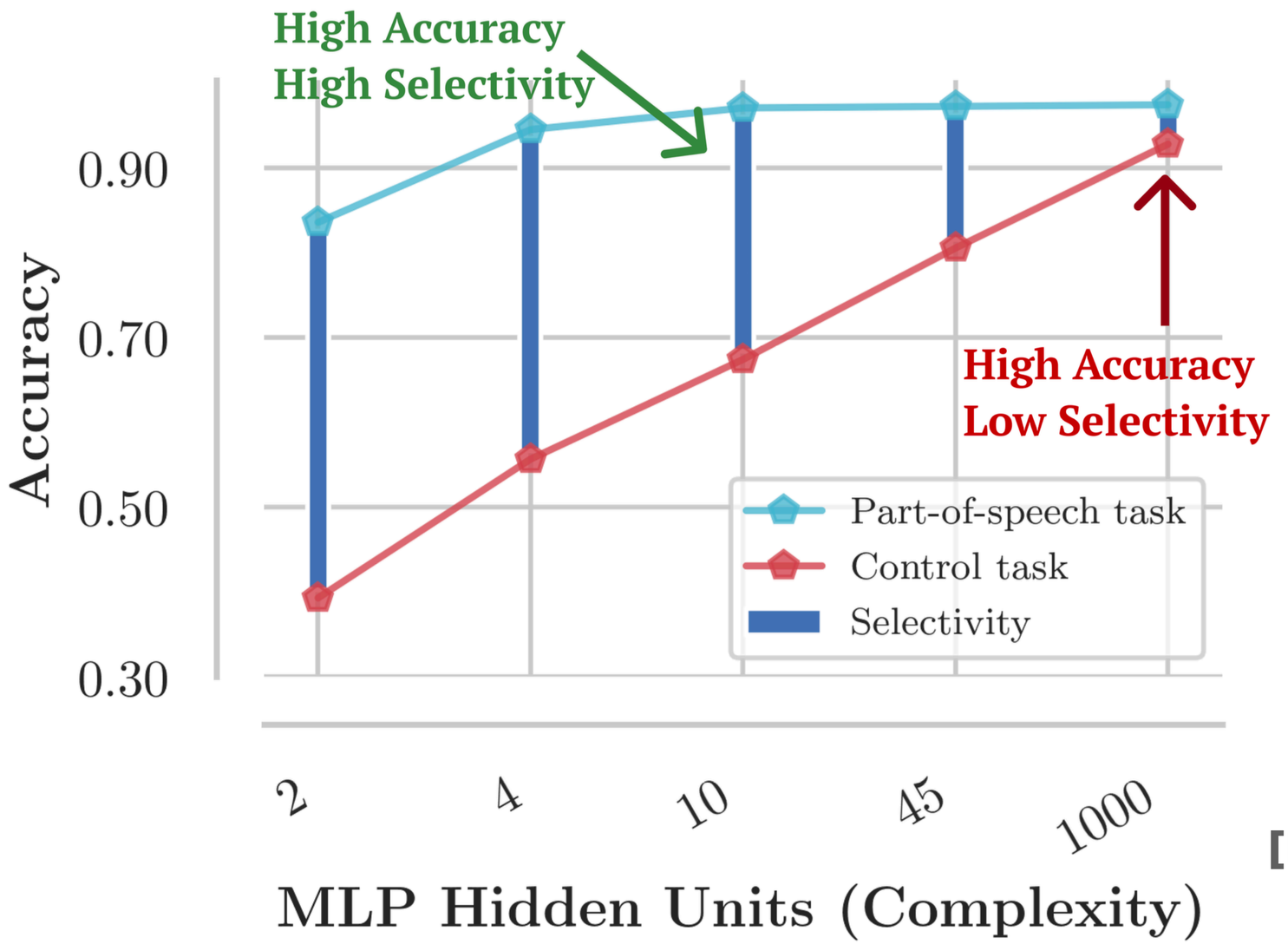
The difference between probe performance on the real task vs. the control task



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]**

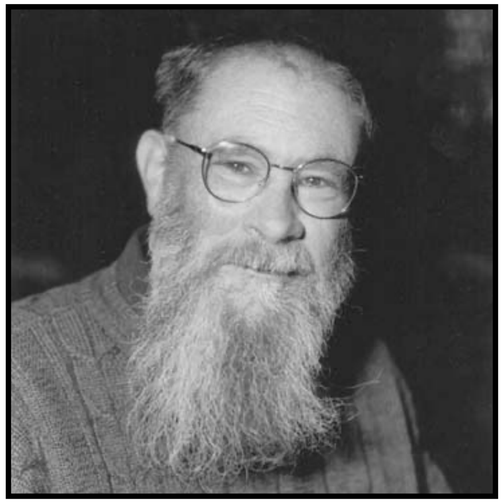
Mechanistic Interpretability

- Probing works by *correlating* model representations to human concepts.

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

What is 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.

Causal Mediation Analysis

Theory



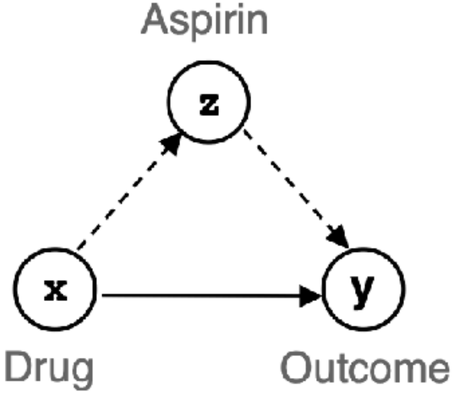
Administer drug

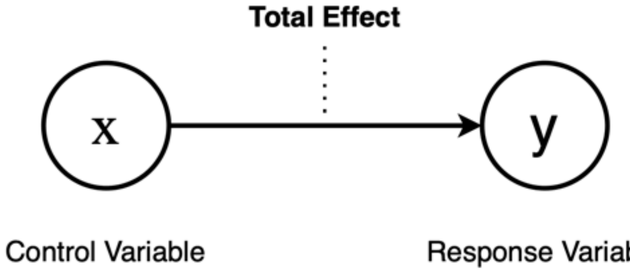


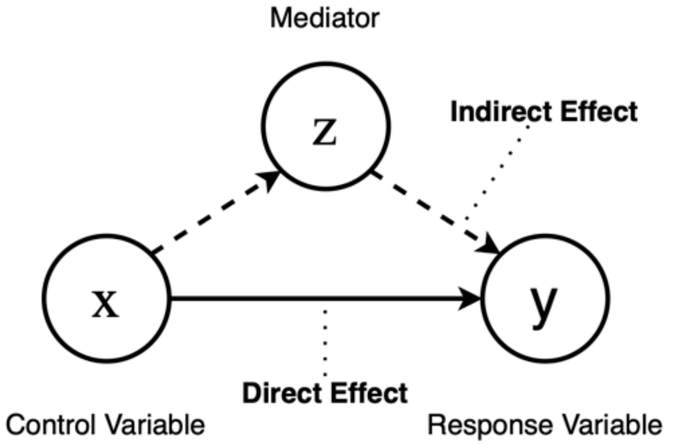
Drug causes headache, aspirin taking



Drug + aspirin → health outcome

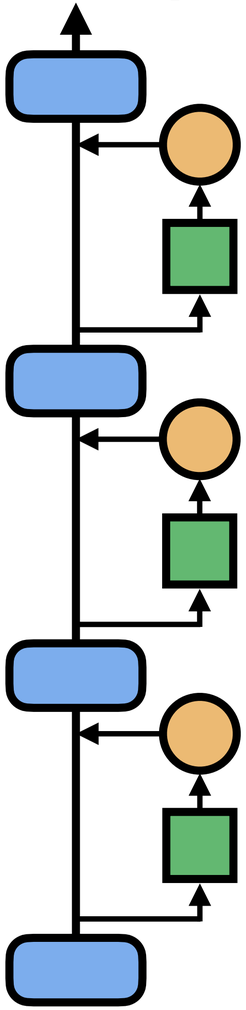


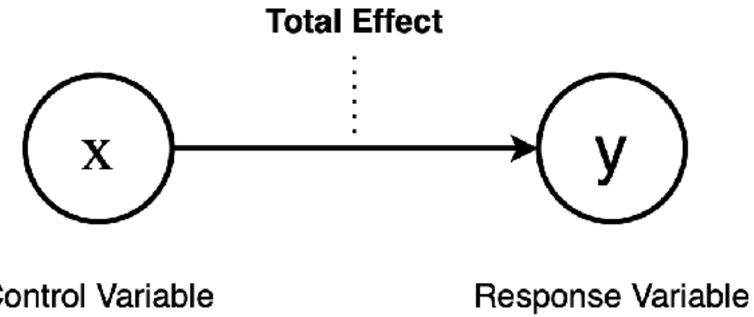




Causal Mediation Analysis

y = p(he) − p(she)

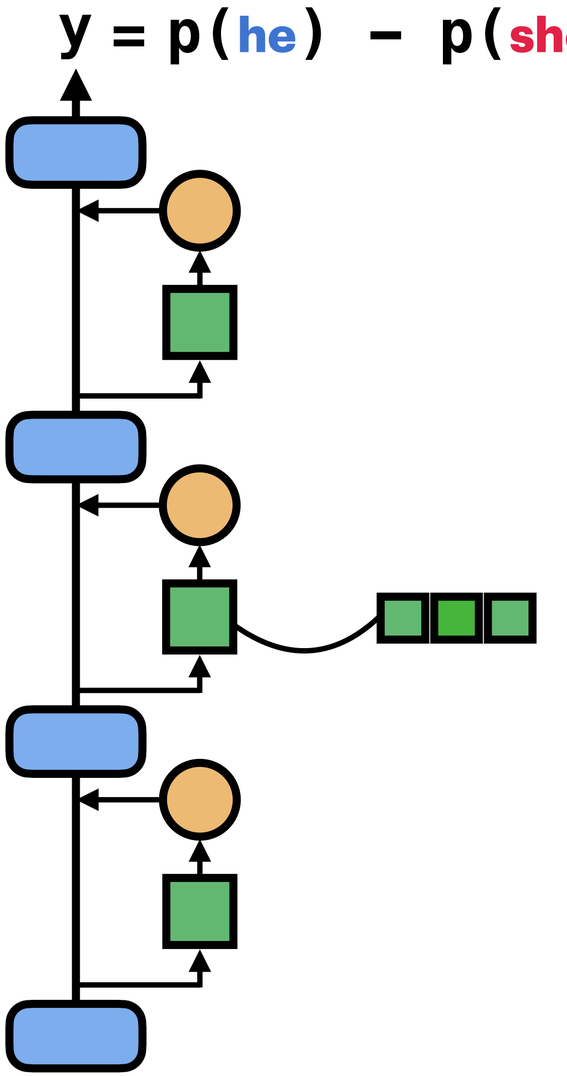




attention MLP
residual vector

x = The doctor administered the treatment because

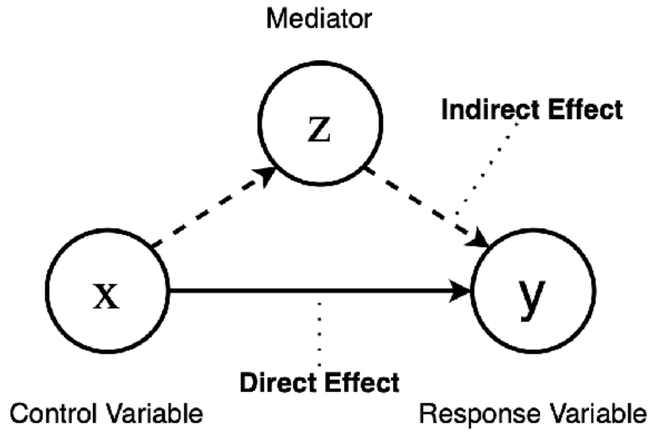
Causal Mediation Analysis





x = The doctor administered the treatment because

Causal Mediation Analysis

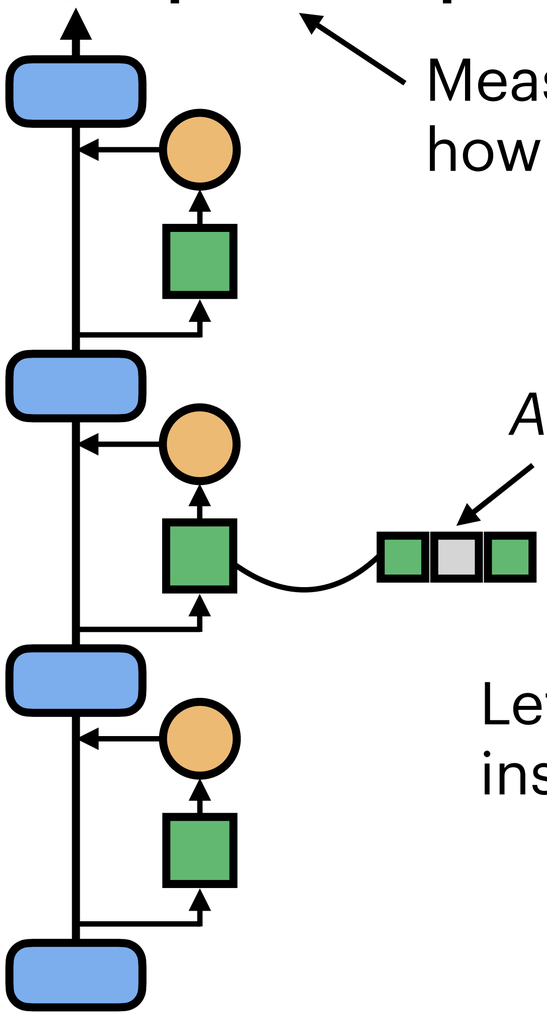


y = p(he) – p(she)

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

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

Ablate this head

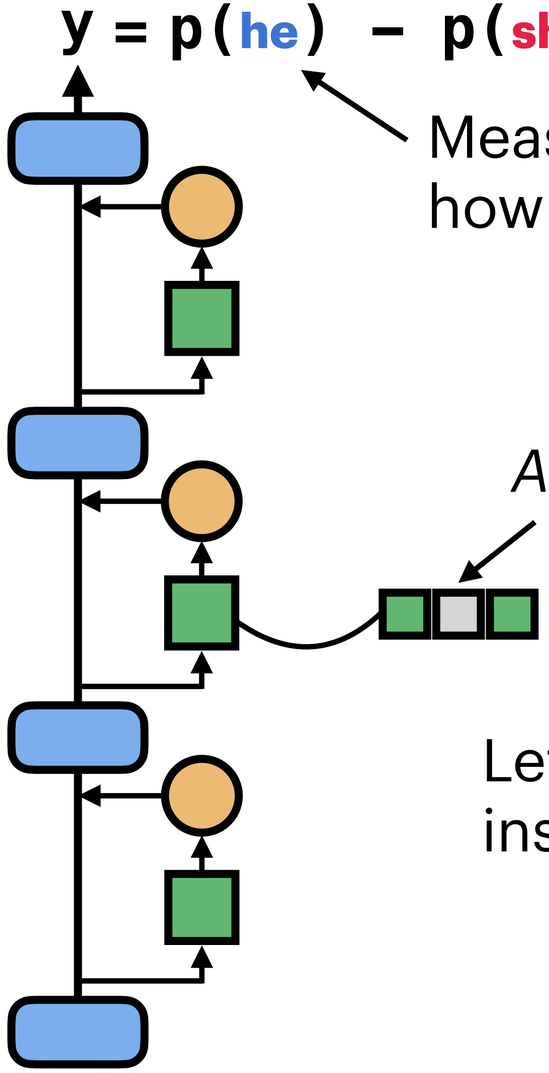


Let's repeat this for each head, and then inspect the ones that change **y** the most.



x = The doctor administered the treatment because

Causal Mediation Analysis



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

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

Ablate this head: set it to 0?

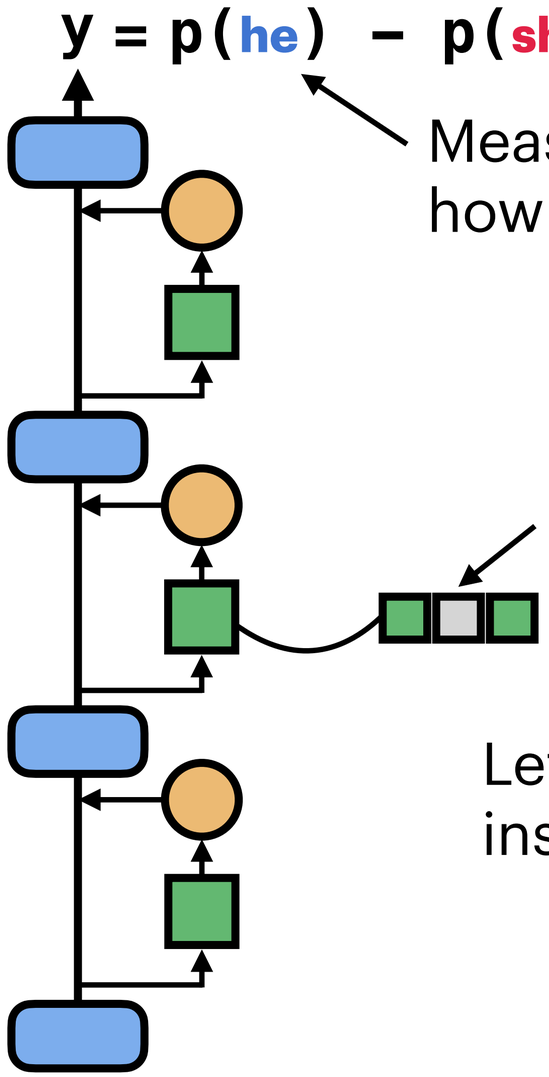
Problem: 0 doesn't have any inherent meaning

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

attention MLP
residual vector

x = The doctor administered the treatment because

Causal Mediation Analysis



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

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

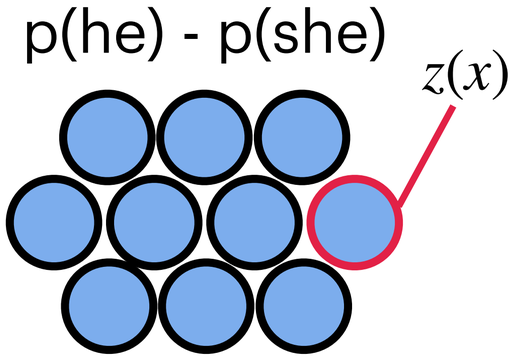
Ablate this head: set it to its *mean value* on some dataset

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

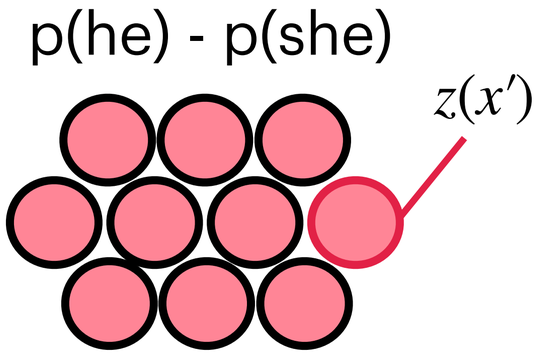
attention MLP
residual vector

x = The doctor administered the treatment because

Example: Causal Mediation Analysis



x = The doctor...



x' = The nurse...

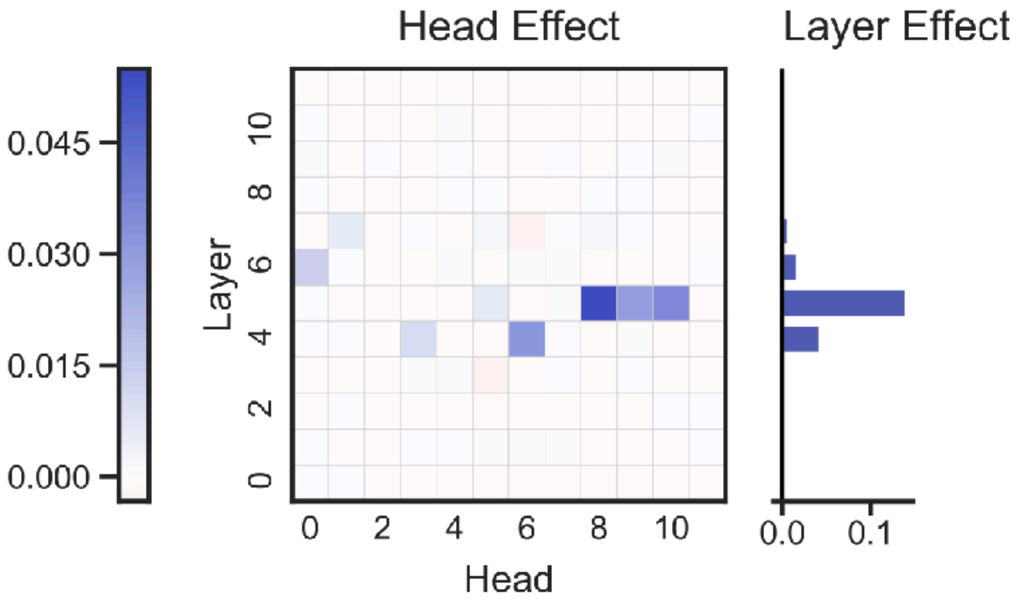
Run	Input	Neuron z	y
1	x	$z(x)$	2
2	x'	$z(x')$	6
3	x	$z(x')$	5
4	x'	$z(x)$	4

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

What is the **indirect effect** of setting z to $z(x')$ given x?

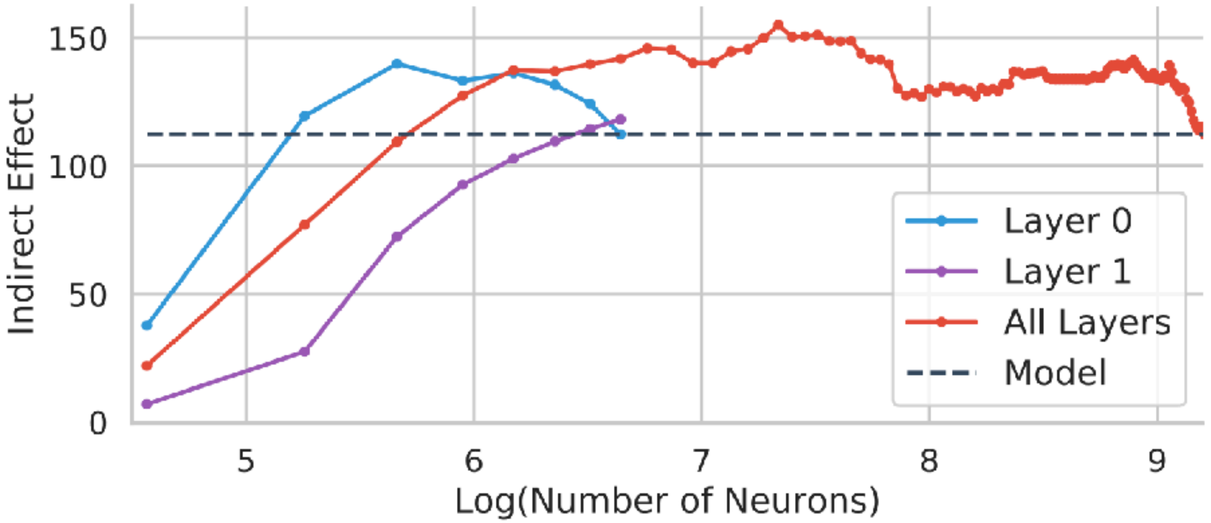
$\text{IE} = y(x \,|\, \text{do}(z = z(x'))) - y(x) = 5 - 2 = 3$

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



Vig et al. [2020] use this method to find gender bias heads and neurons in LMs.

They also measure the effect of multiple neurons at once:



Note: the IE declines on the right side of this plot. What does this tell us?

That some components have *negative* IE, and thus counteract the influence of the components with positive IE.

Can we use interpretability to fix models?

- Understanding where models implement a certain behavior is nice...
- ...but ideally, we'd like to be able to *fix* models.

- 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

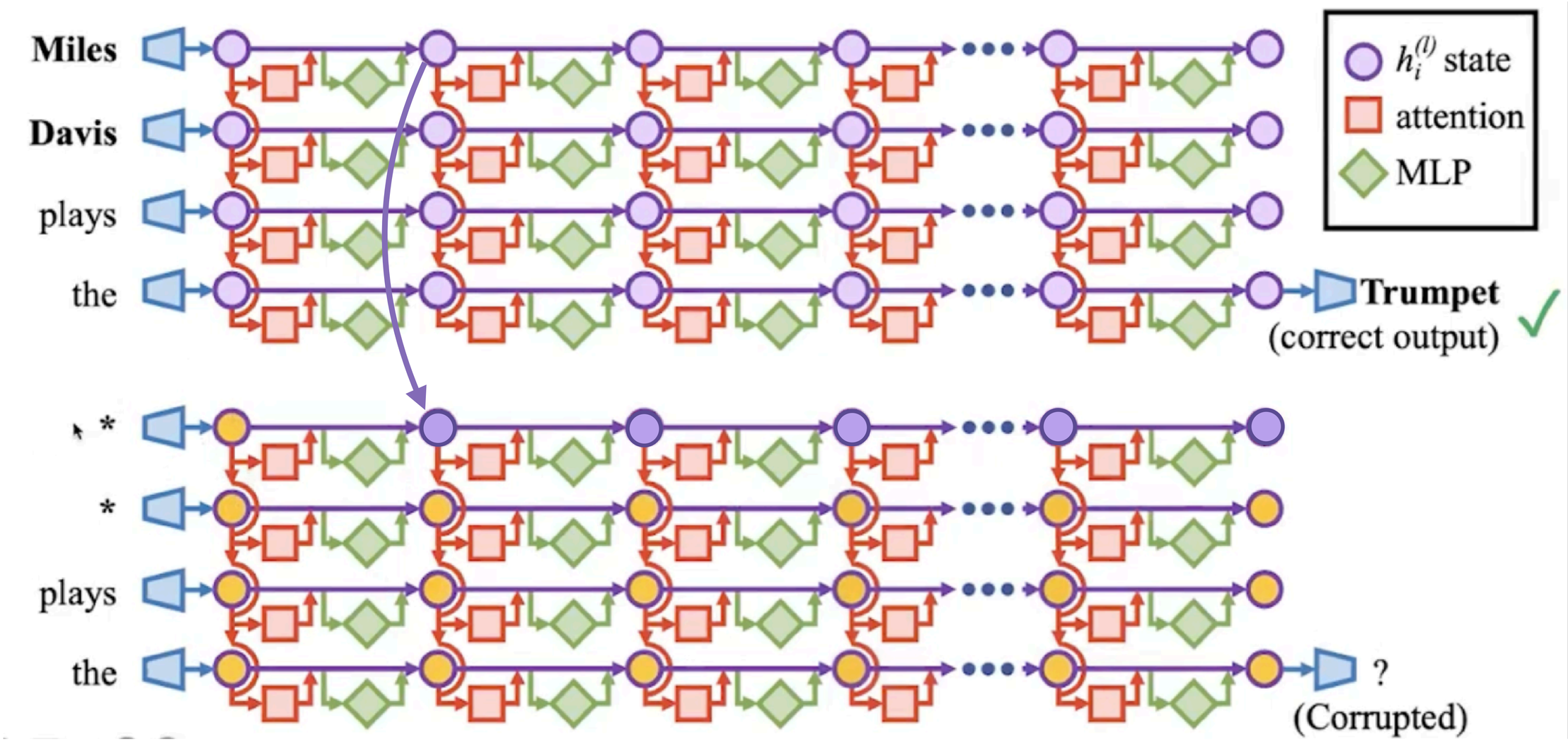
Can we precisely **localize** where a behavior is being computed in a neural network?

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

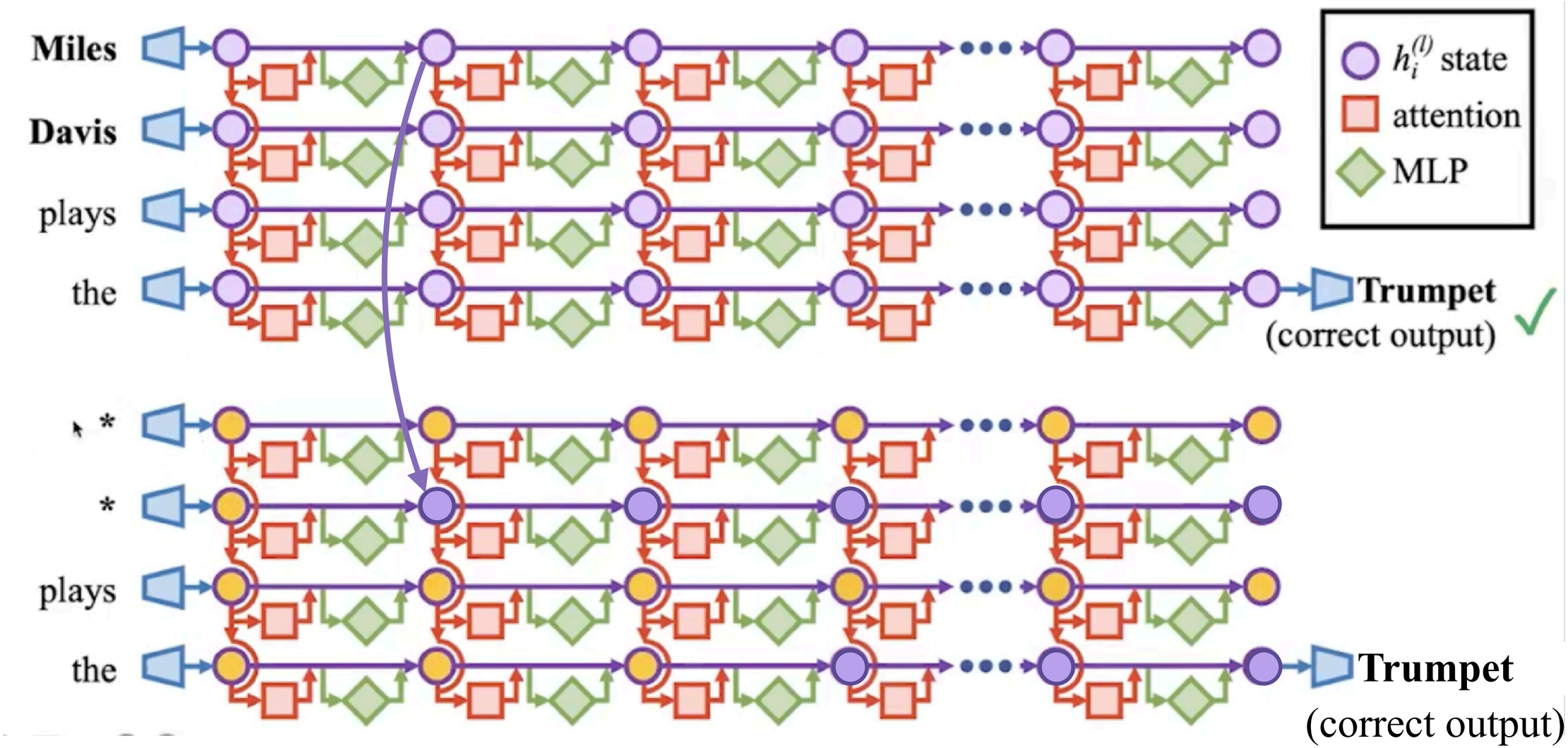
Yes! We'll go over one method today, and another next time:

1. *Model editing* with ROME
2. *Steering* with steering vectors or manifolds

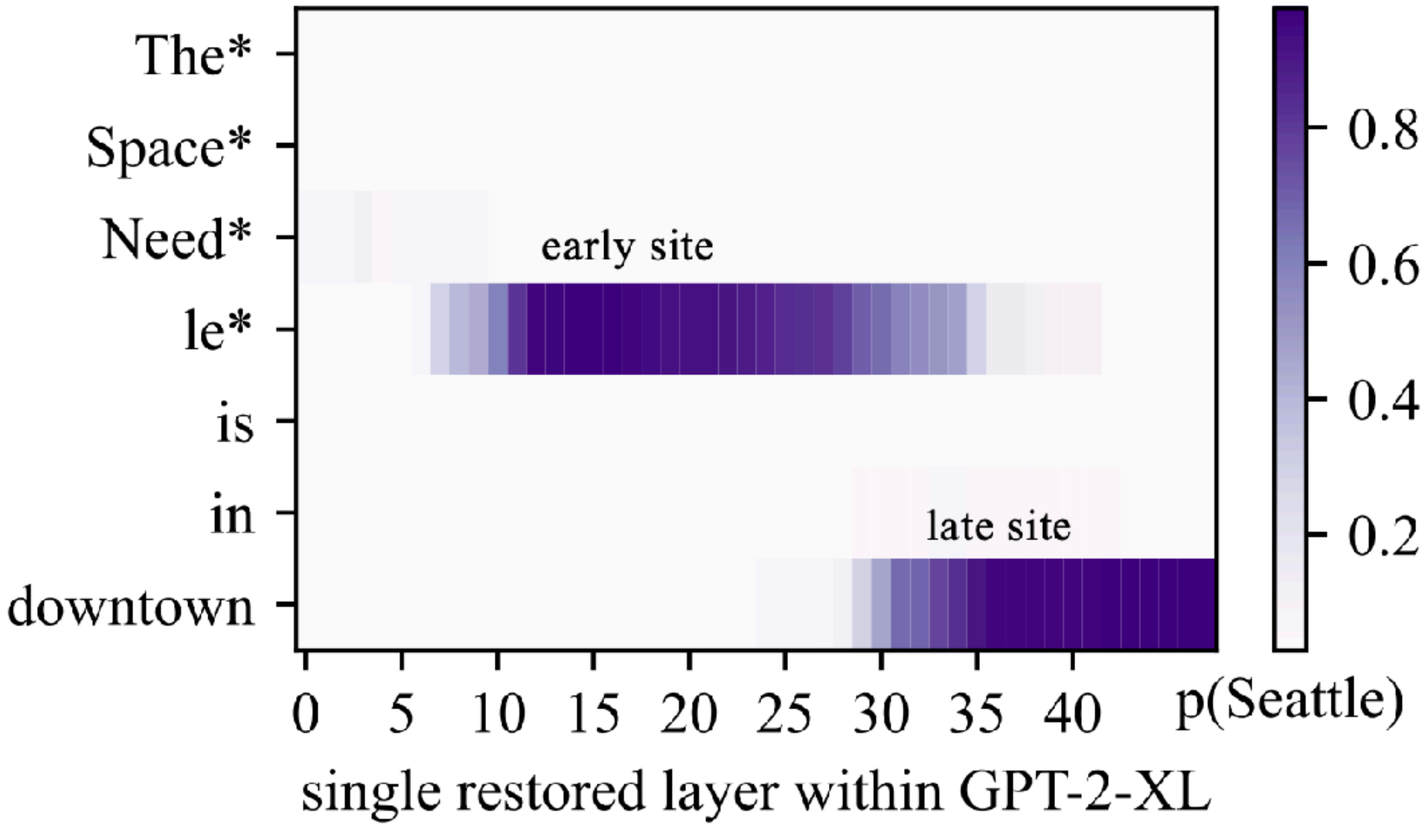
Locating Knowledge in LMs



Locating Knowledge in LMs



Causal Tracing



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

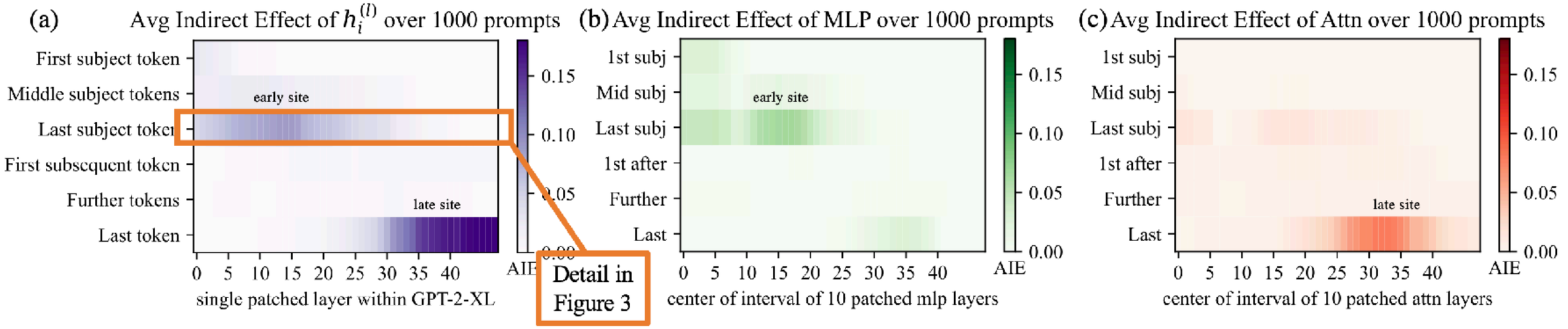
The **early site**

The **late site**

What's happening in these sites?

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

Causal Tracing



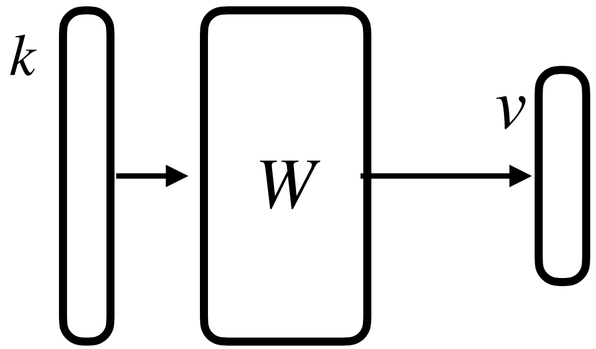
- Knowledge is encoded quite consistently across facts:
  - Two sites
  - Last subject token
  - Mid-layer MLP and late-layer attention modules do a lot of the work

Associative Memory View of a Neural Layer

- Why are we able to find knowledge like this?
- Some have posited that we 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 a representation of the query (e.g., subject and relation), and v_i is the completion of the fact.



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