




Sparse Feature Circuits

Discovering and Editing Interpretable Causal Graphs in Language Models

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, **Aaron Mueller**

2025 International Conference on Learning Representations (ICLR)
26 April 2025

Interpretability

For a model to generalize, it must achieve right answers *for the right reasons*.

Interpretability

For a model to generalize, it must achieve right answers *for the right reasons*.

How do neural networks (NNs) perform particular behaviors?

Why do they behave in certain ways on certain inputs?

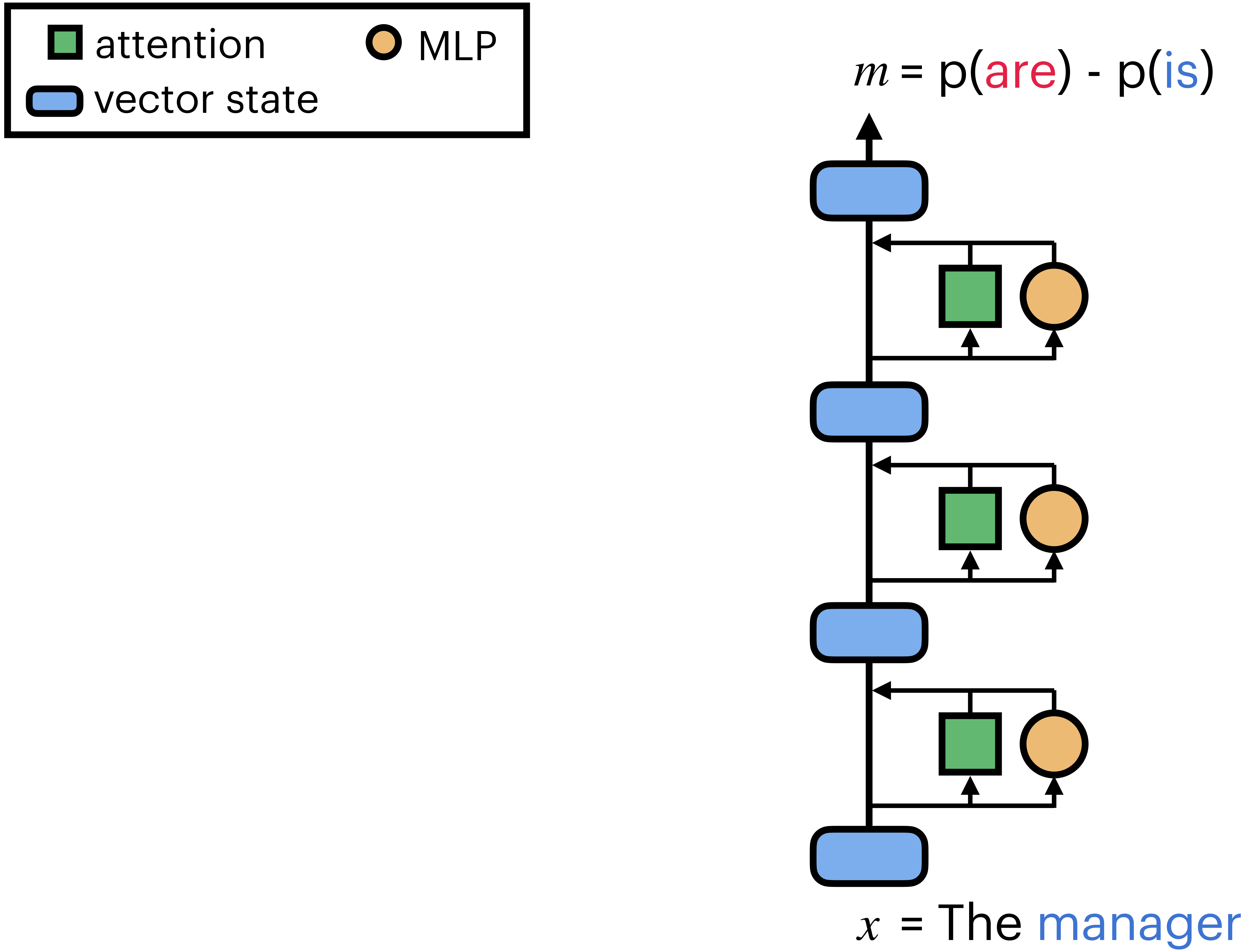
Interpretability

For a model to generalize, it must achieve right answers *for the right reasons*.

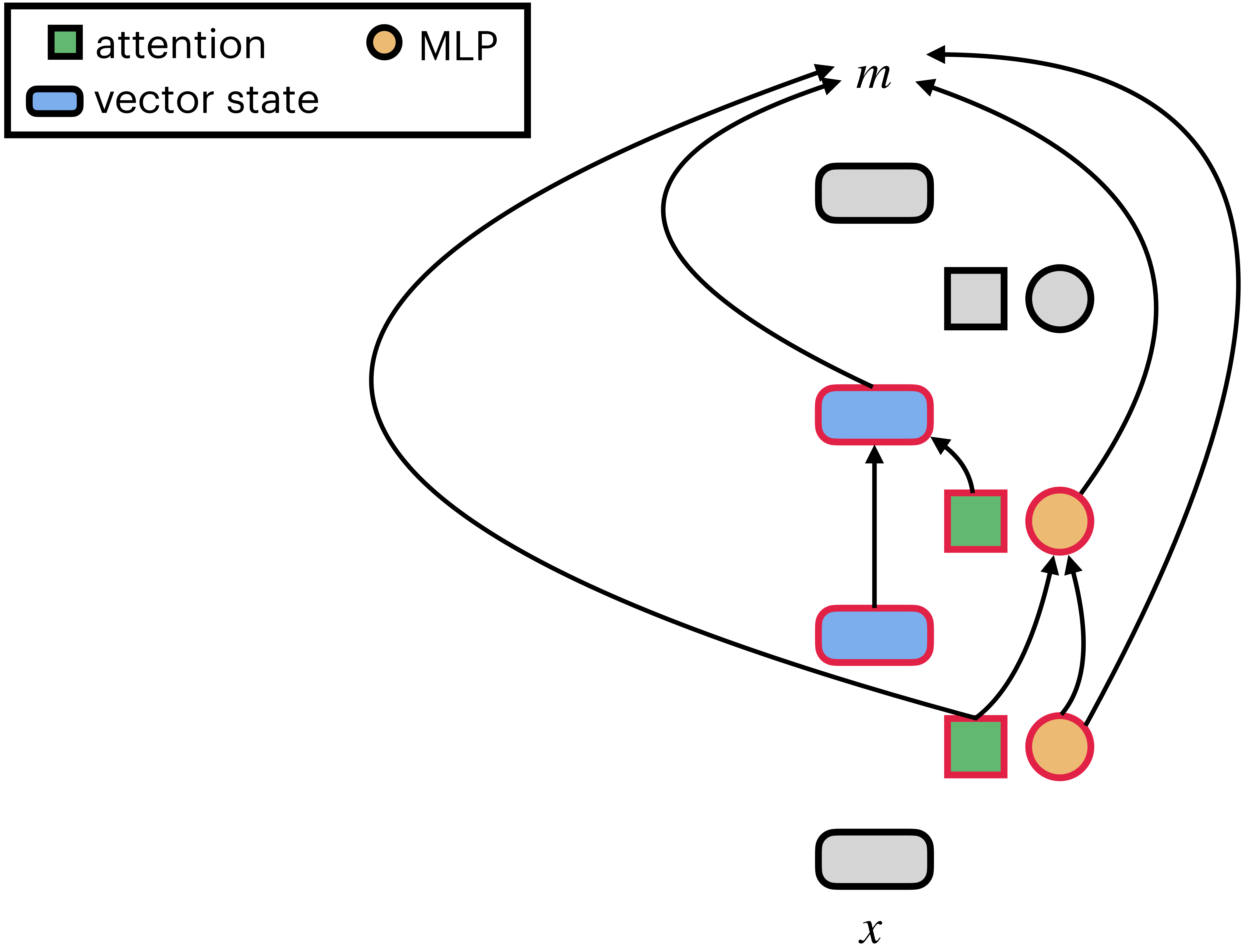
How do neural networks (NNs) perform particular behaviors?

Why do they behave in certain ways on certain inputs?

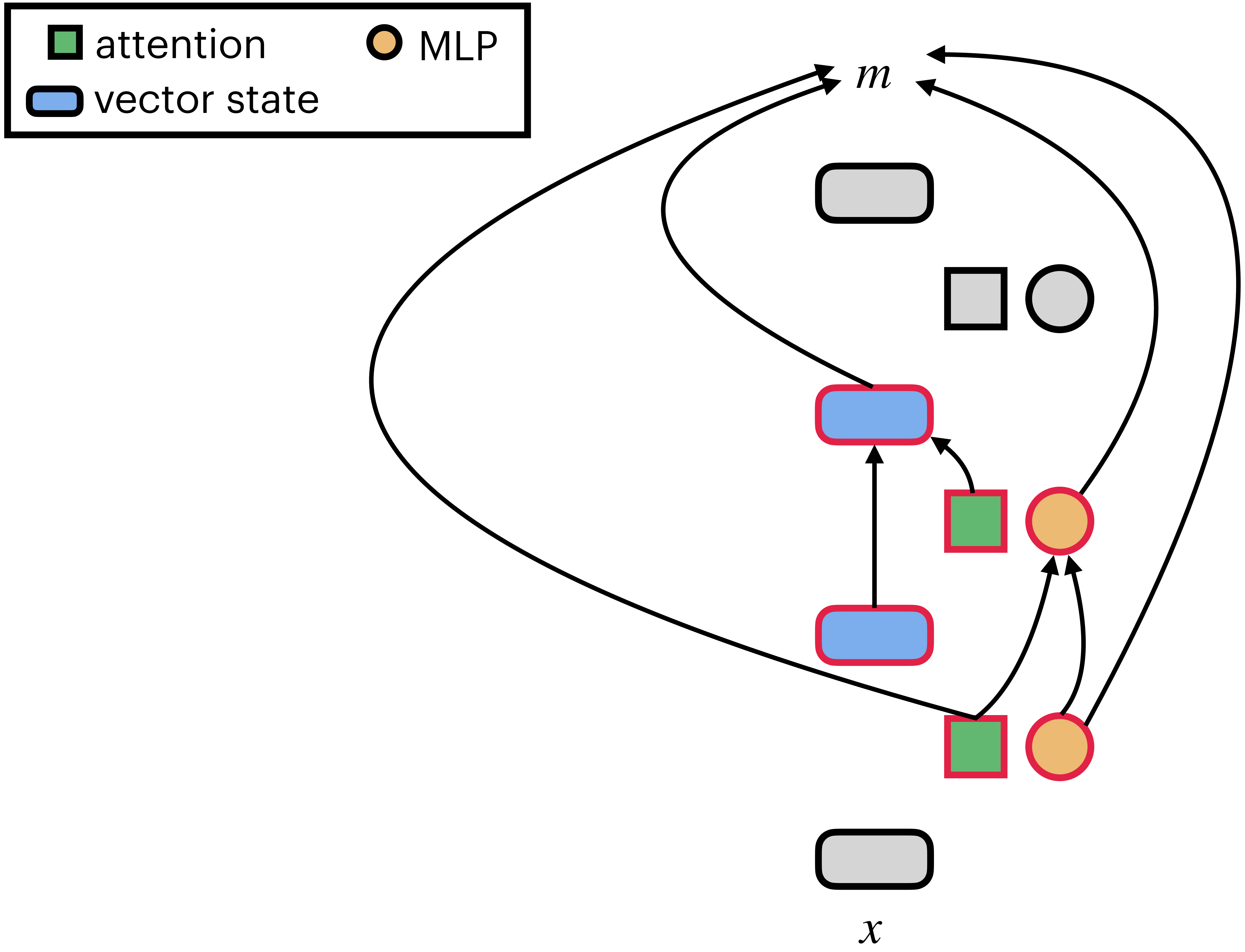
*How can we locate and understand **unanticipated mechanisms**?*



Given a neural network, we want to know which components contribute most to the model’s behavior.

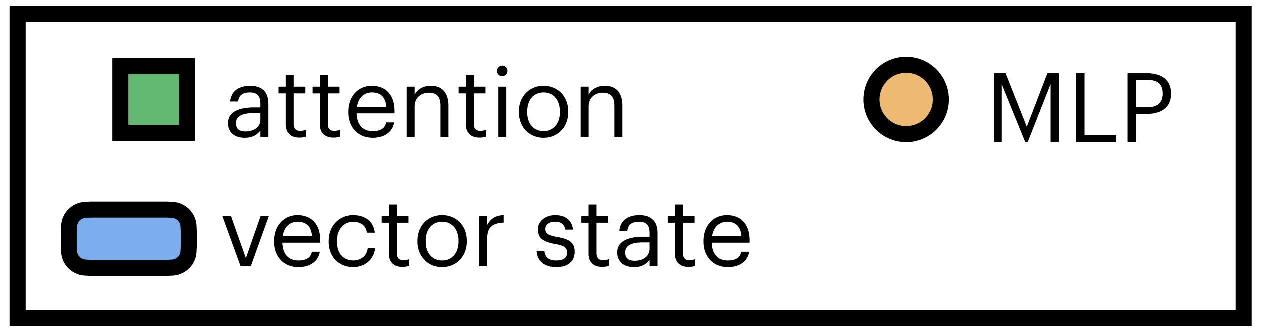


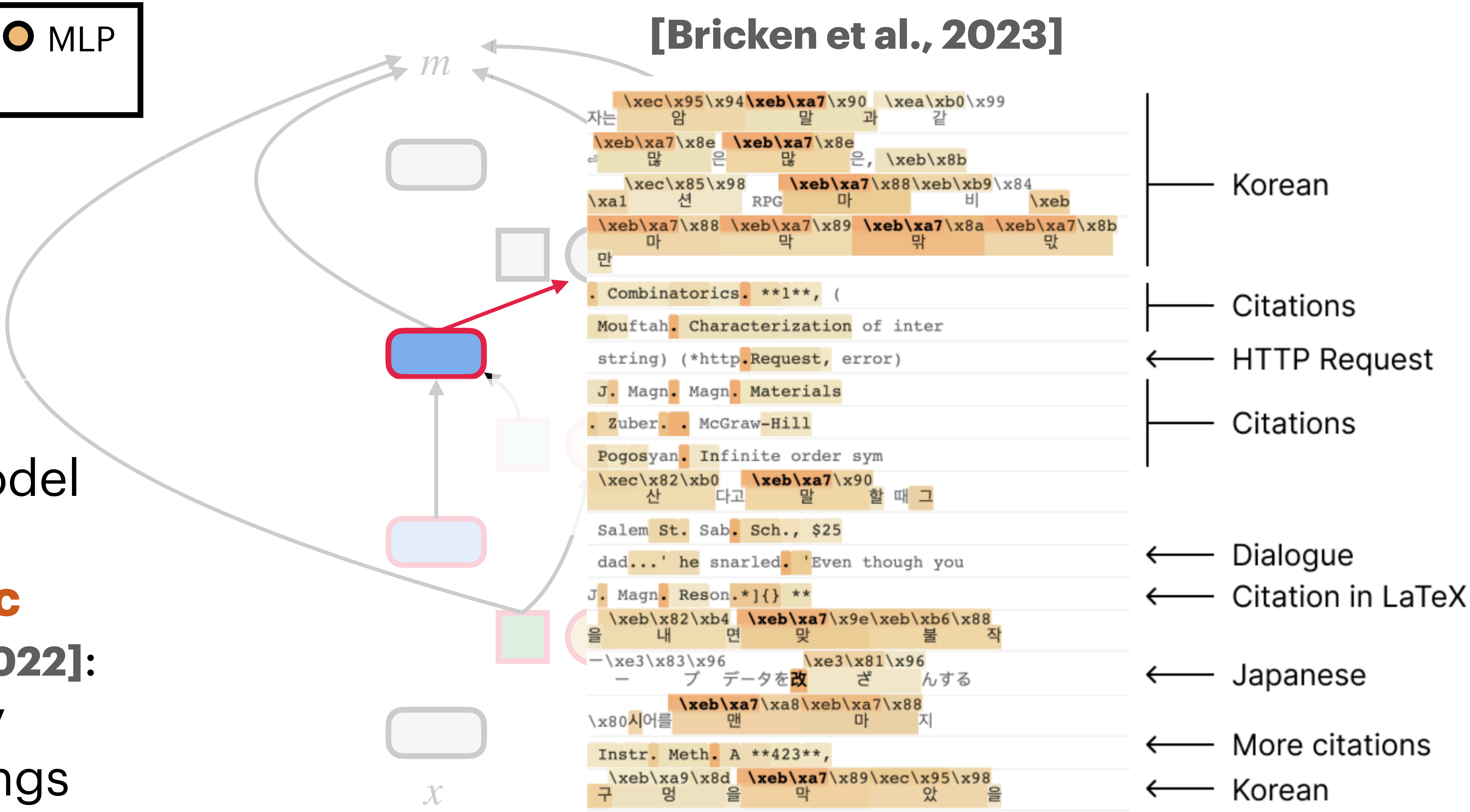
We have a **circuit**!



We have a **circuit**!

...Now what?





Language model neurons are **polysemantic** **[Elhage et al., 2022]**: they do many unrelated things simultaneously.

Nelson Elhage et al. (2022). "Toy Models of Superposition." *Transformer Circuits Thread.*
Trenton Bricken et al. (2023). "Towards Monosemanticity: Decomposing Language Models with Dictionary Learning." *Transformer Circuits Thread.*

Maybe this is the wrong abstraction. We need to decompose **polysemantic** representations into **monosemantic** units.

Language model neurons are **polysemantic [Elhage et al., 2022]**: they do many unrelated things simultaneously.

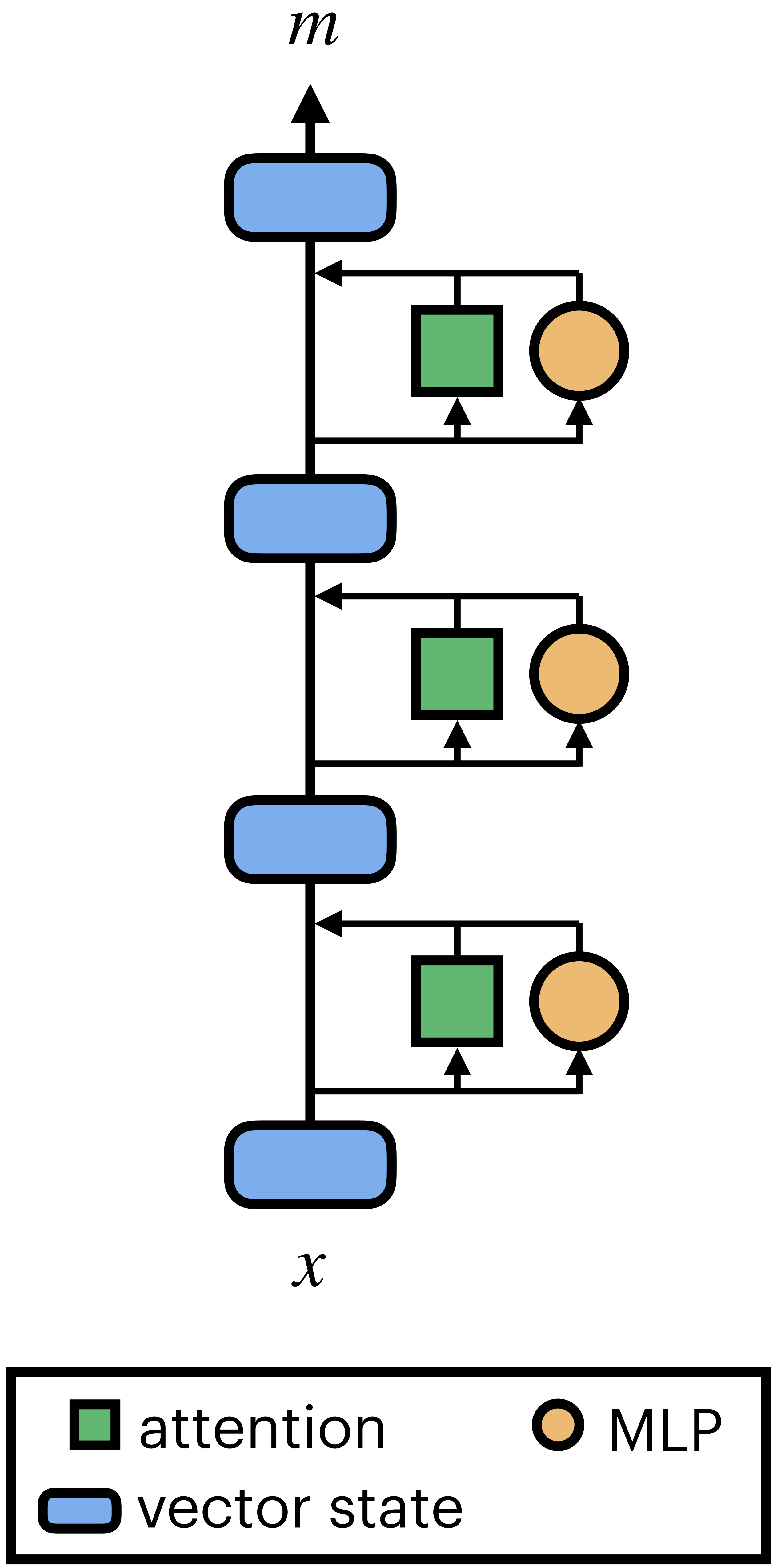
Nelson Elhage et al. (2022). "Toy Models of Superposition." *Transformer Circuits Thread.*

Trenton Bricken et al. (2023). "Towards Monosemanticity: Decomposing Language Models with Dictionary Learning." *Transformer Circuits Thread.*

Maybe this is the wrong abstraction. We need to decompose **polysemantic** representations into **monosemantic** units.

Goal: a circuit where each node is *human-interpretable* and *editable*

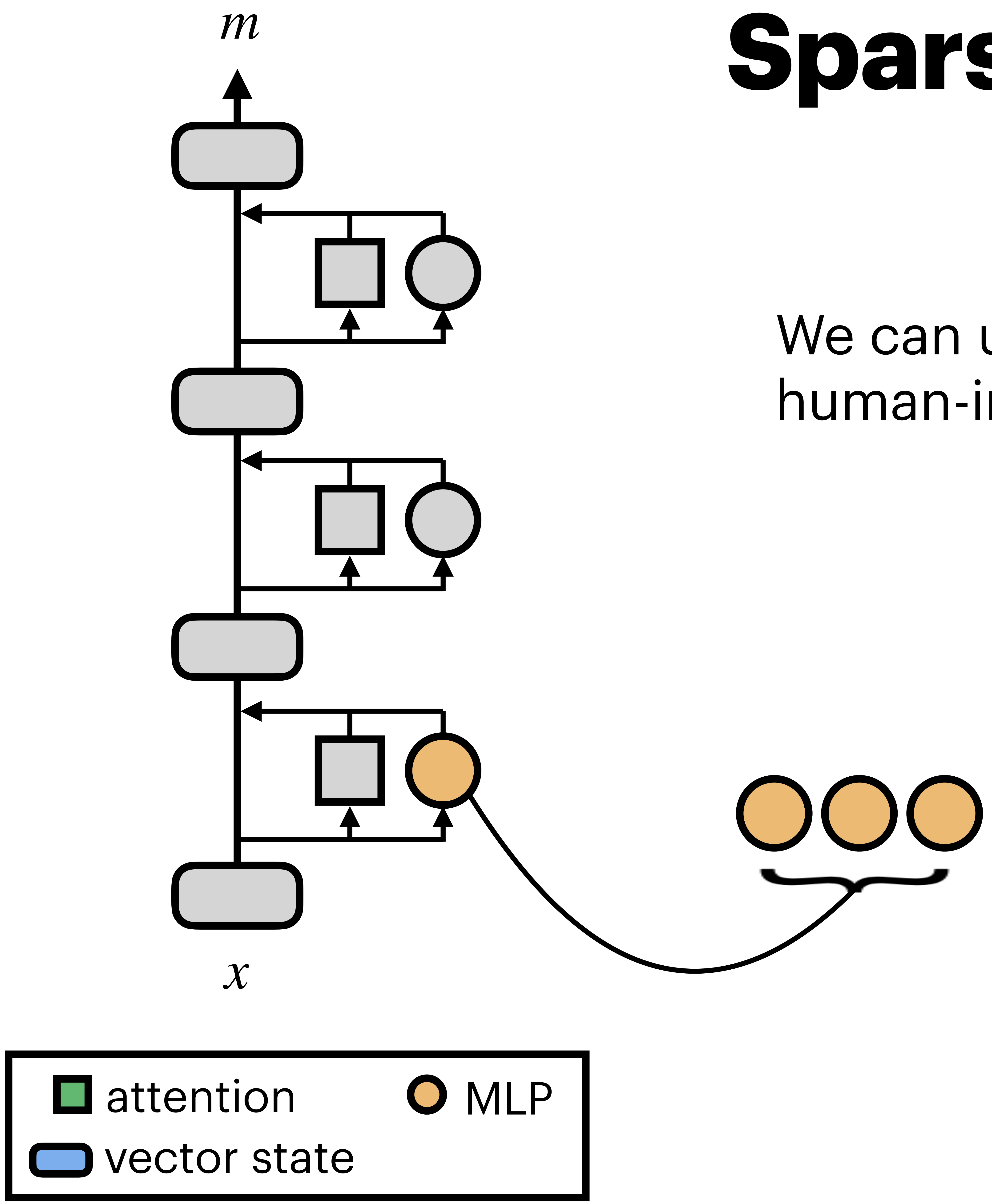
Nelson Elhage et al. (2022). "Toy Models of Superposition." *Transformer Circuits Thread.*
Trenton Bricken et al. (2023). "Towards Monosemanticity: Decomposing Language Models with Dictionary Learning." *Transformer Circuits Thread.*



Sparse Features

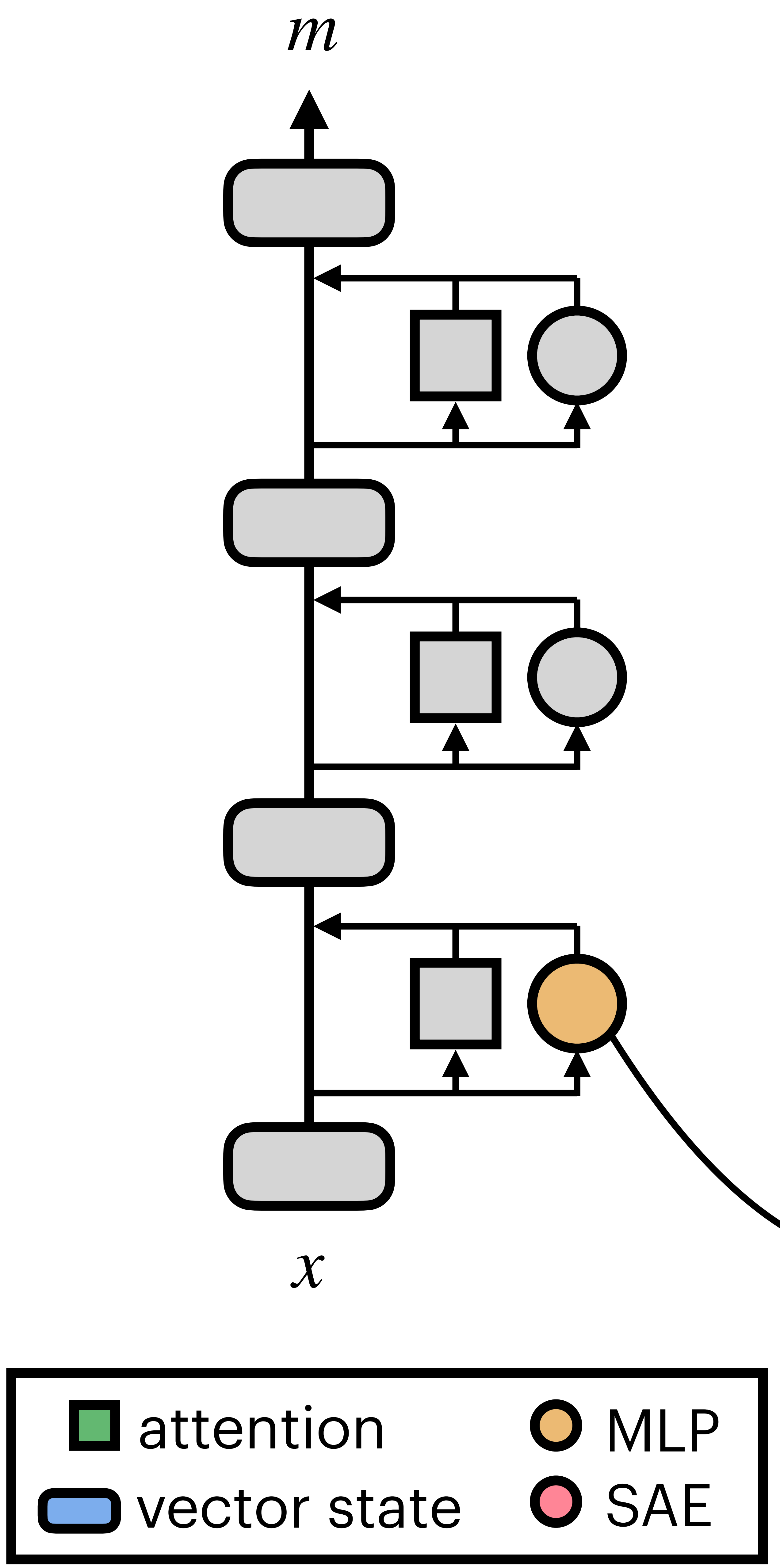
We can use **sparse autoencoders** (SAEs) to disentangle human-interpretable **features** from model components

Sparse Features

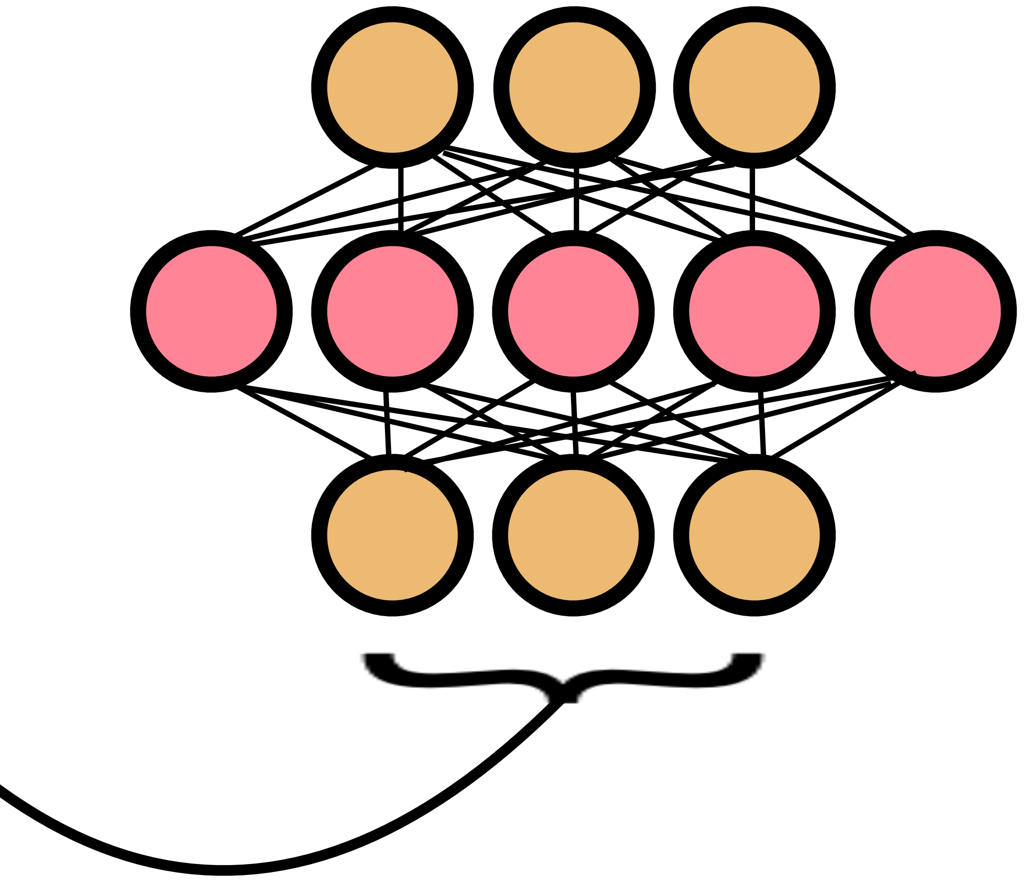


We can use **sparse autoencoders** (SAEs) to disentangle human-interpretable **features** from model components

Sparse Features



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

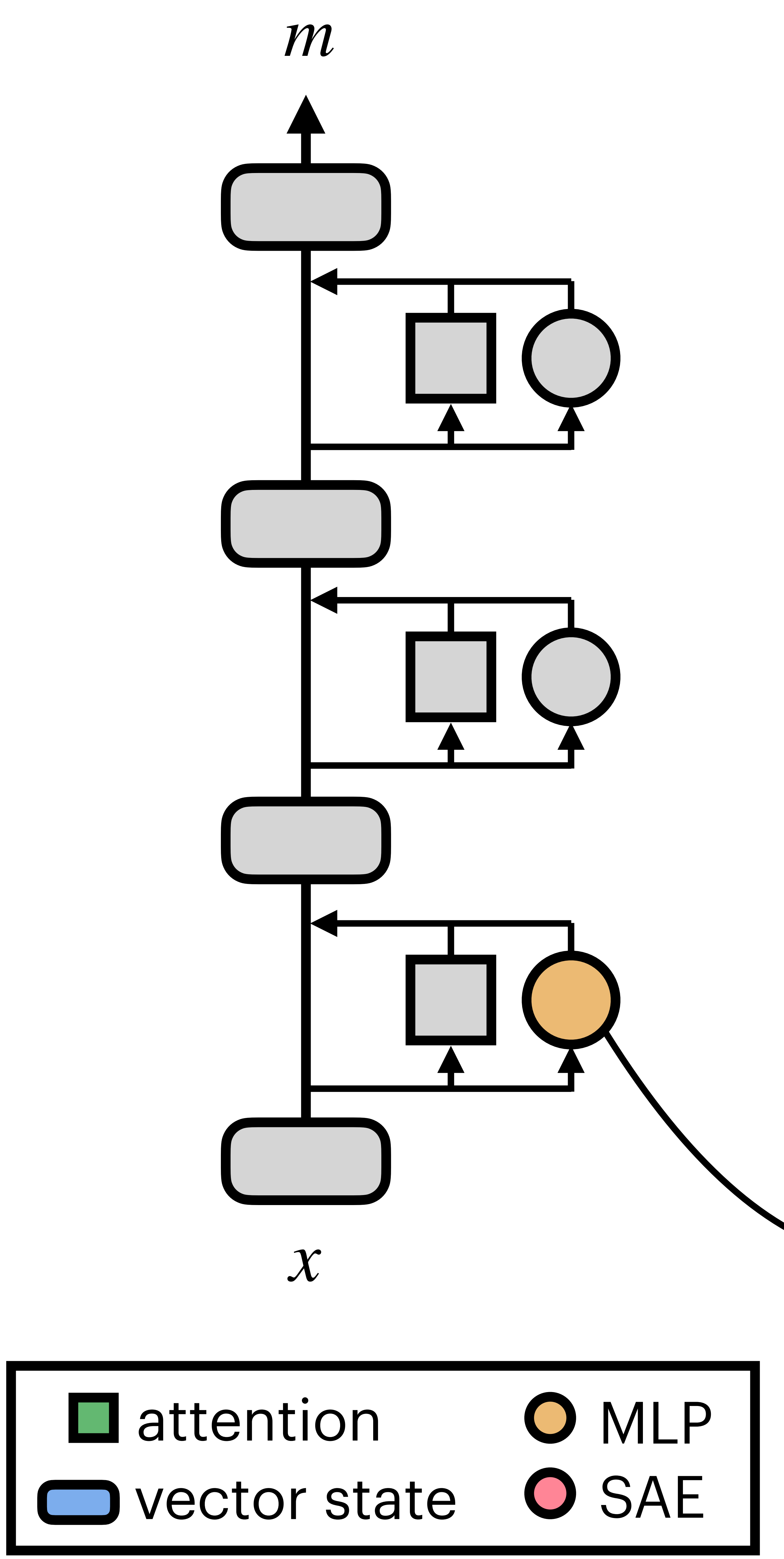


$$\hat{\mathbf{x}} = W_d\mathbf{f} + \mathbf{b}_d$$

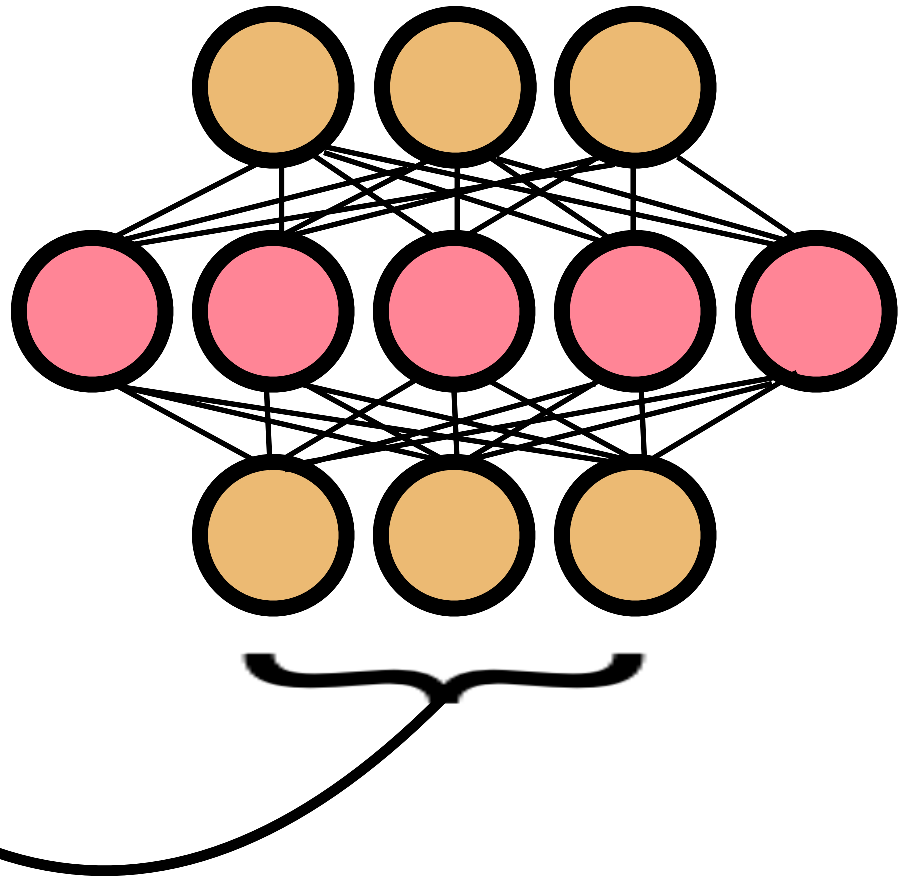
$$\mathbf{f} = \mathrm{ReLU}(W_e(\mathbf{x} - \mathbf{b}_d) + \mathbf{b}_e)$$

$$\mathbf{x}$$

Sparse Features



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



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

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

$$\mathbf{x}$$

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

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

Sparse Features

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

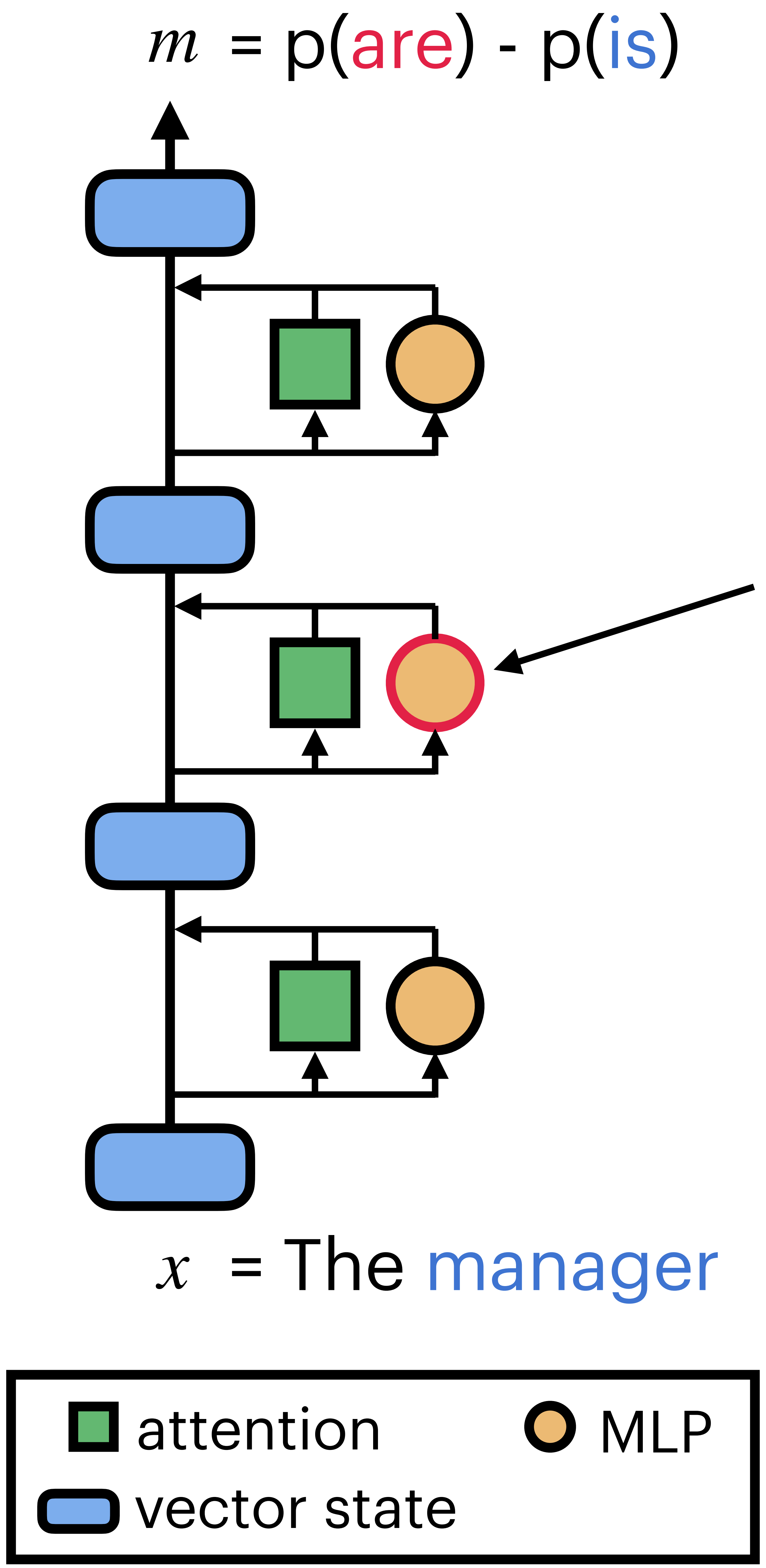
Passages related to academia, research

Language Models and SAEs

- **Models:** Pythia (70M), Gemma 2 (2B)
- **SAEs:** GemmaScope **[Lieberum et al, 2024]**, or trained by us on model activations given documents from The Pile
- SAE features are interpreted using activations and logits from The Pile

Tom Lieberum et al. (2024). "Gemma Scope: Open Sparse Autoencoders Everywhere All at Once On Gemma 2." *BlackboxNLP*.

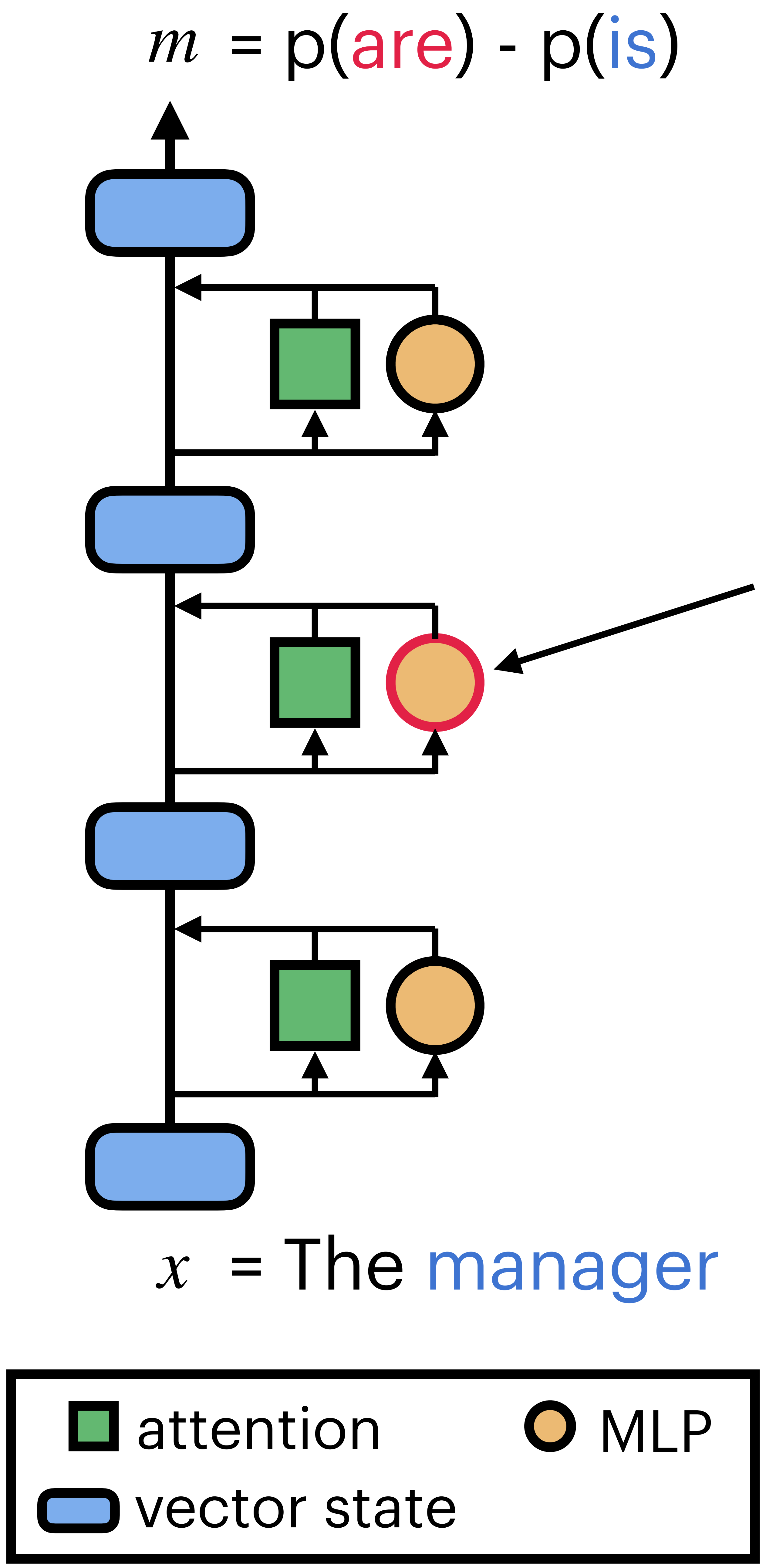
Activation Patching



x' = The ~~manager~~ managers

do(`swap-number`): Set a to what it would have been if the subject in x were the opposite number

Activation Patching



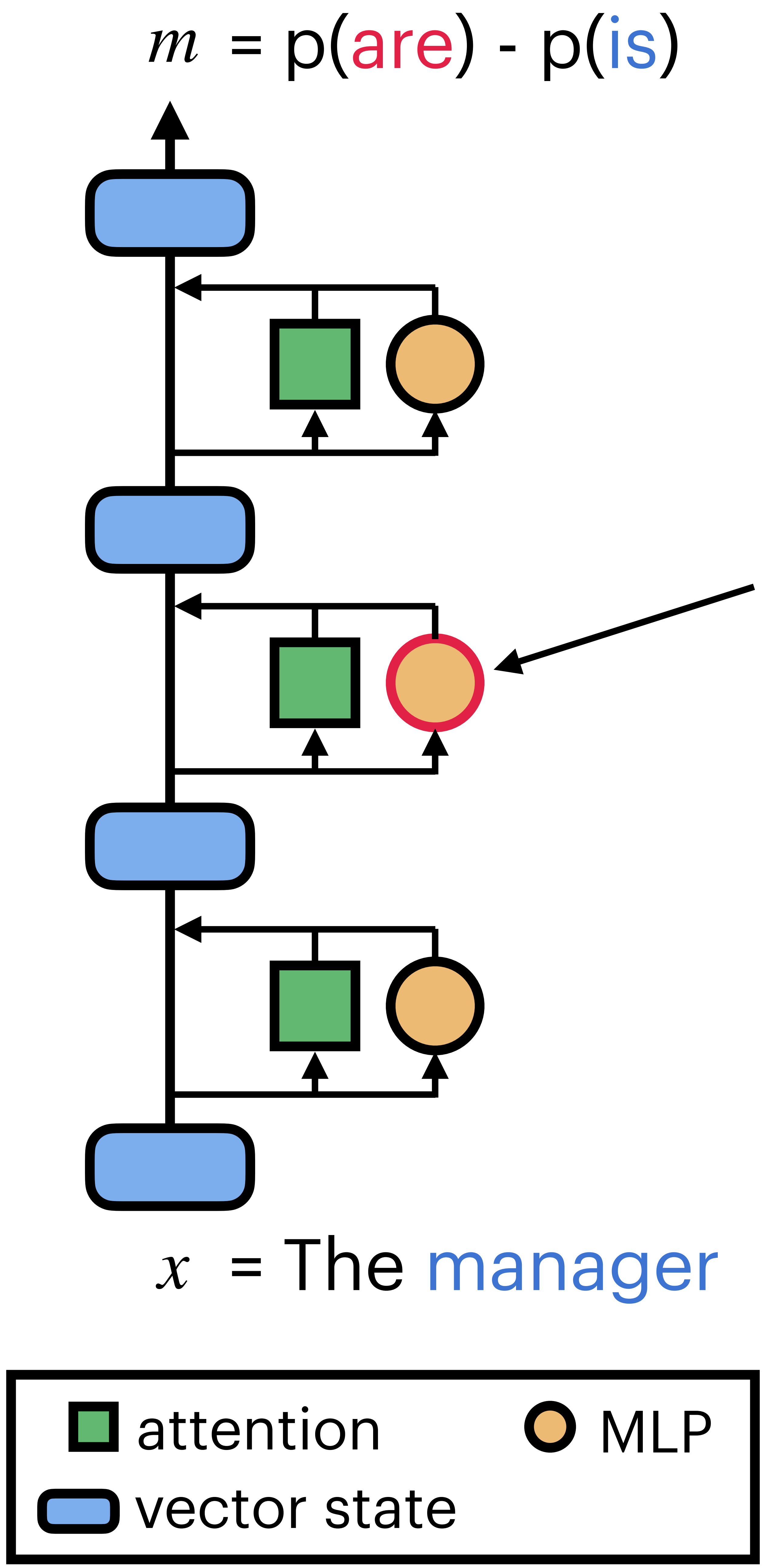
x' = The ~~manager~~ managers

do(`swap–number`): Set a to what it would have been if the subject in x were the opposite number

Indirect effect$(m; a; x, x')$: How much does do(`swap–number`) change m?

$$\mathrm{IE}(m; a; x, x') = m(x, \mathrm{do}(a = a_{x'})) - m(x)$$

Activation Patching



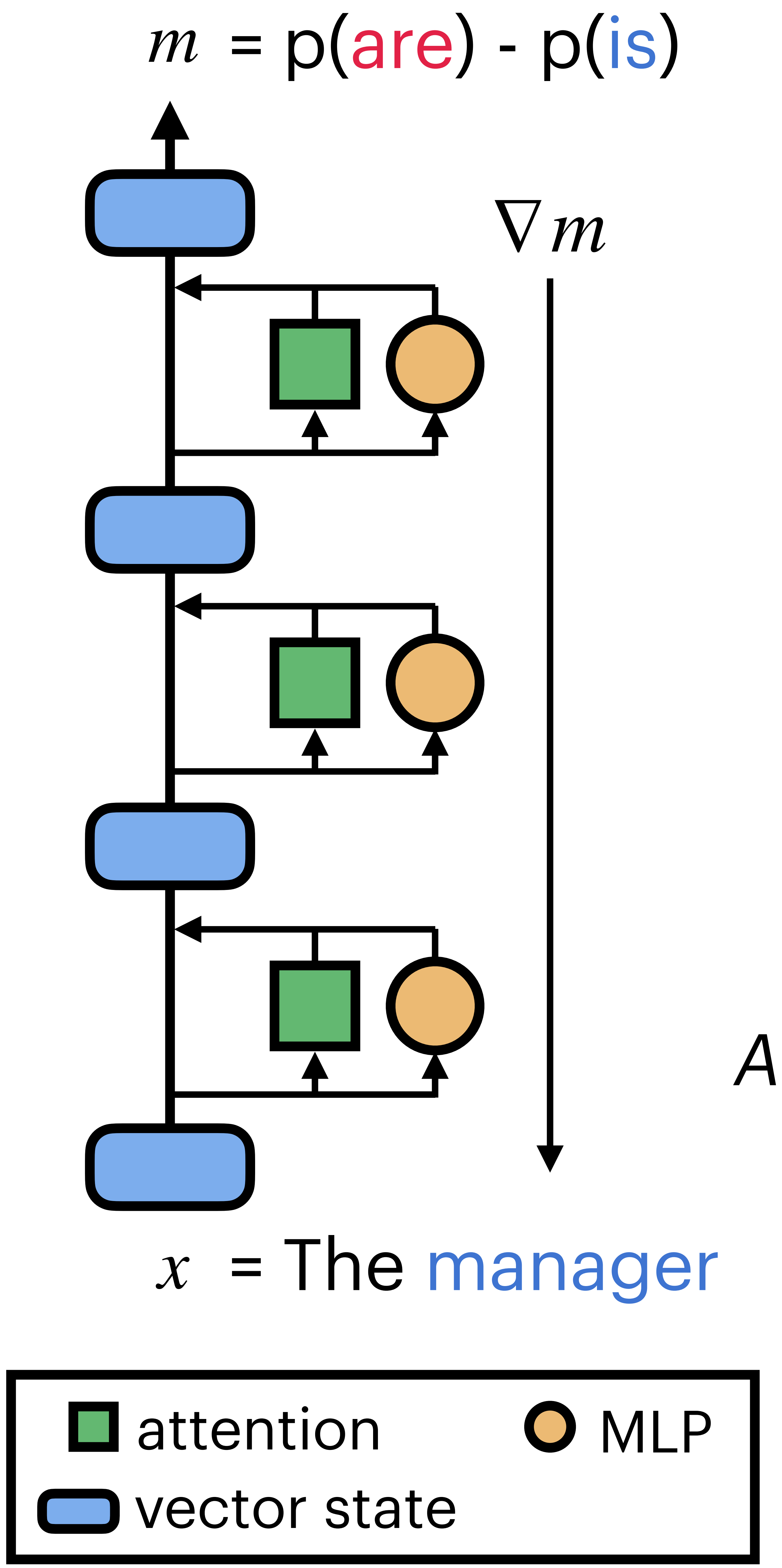
x' = The ~~manager~~ managers

do(`swap-number`): Set a to what it would have been if the subject in x were the opposite number

Indirect effect$(m; a; x, x')$: How much does do(`swap-number`) change m?

$$\mathrm{IE}(m; a; x, x') = m(x, \mathrm{do}(a = a_{x'})) - m(x)$$

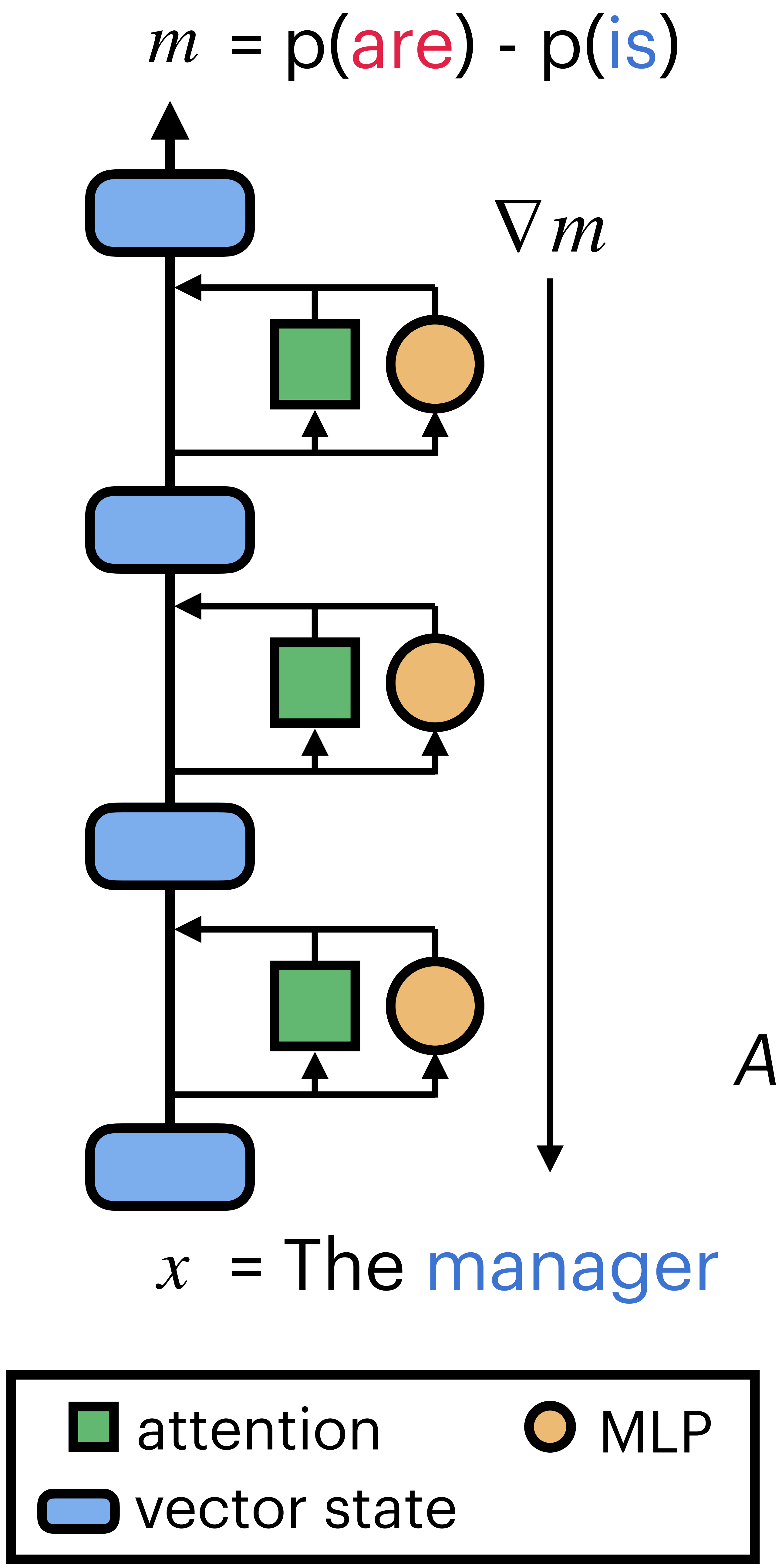
Activation patching requires $O(\mathbf{a})$ forward passes.



Attribution Patching

$$\hat{\text{IE}}(m; a; x, x') = \frac{\partial m}{\partial a}\bigg|_{x} (a_{x'} - a_x)$$

Attribution patching requires $O(1)$ forward and backward passes!

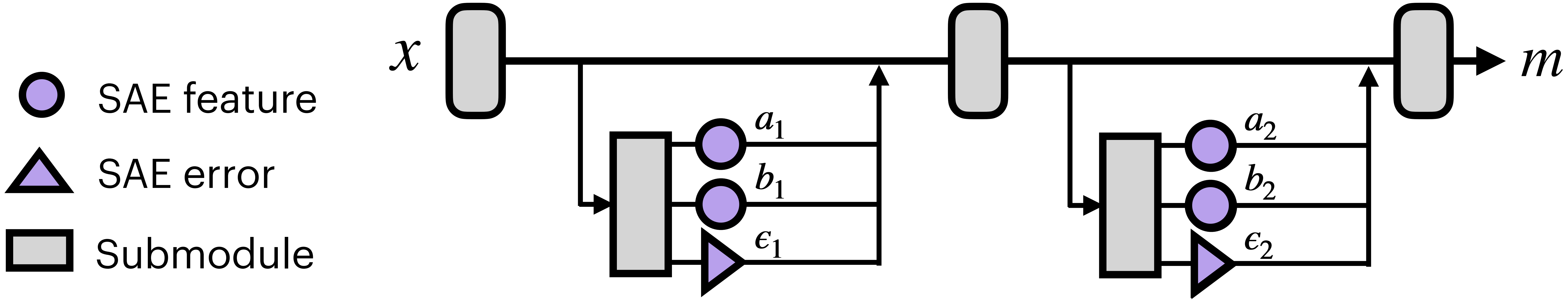


Attribution Patching

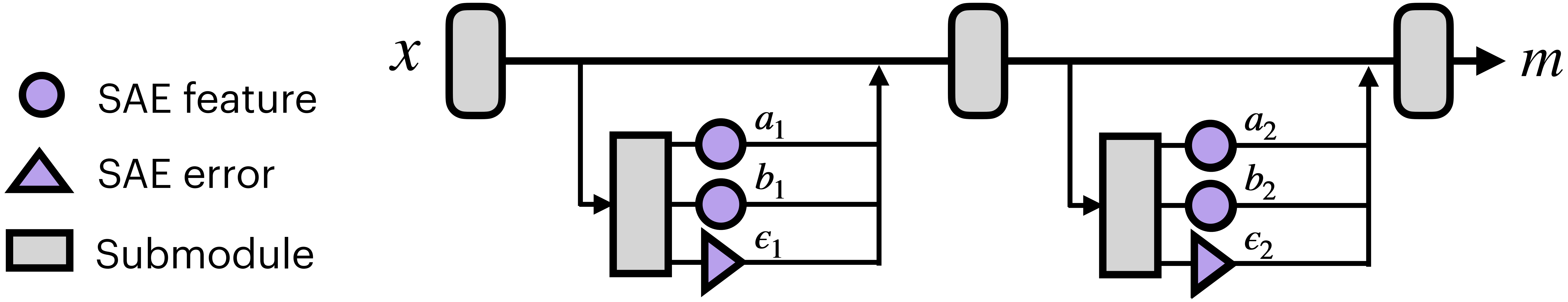
$$\hat{\mathrm{IE}}(m; a; x, x') = \frac{\partial m}{\partial a}\bigg|_x (a_{x'} - a_x)$$

Attribution patching requires $O(1)$ forward and backward passes!

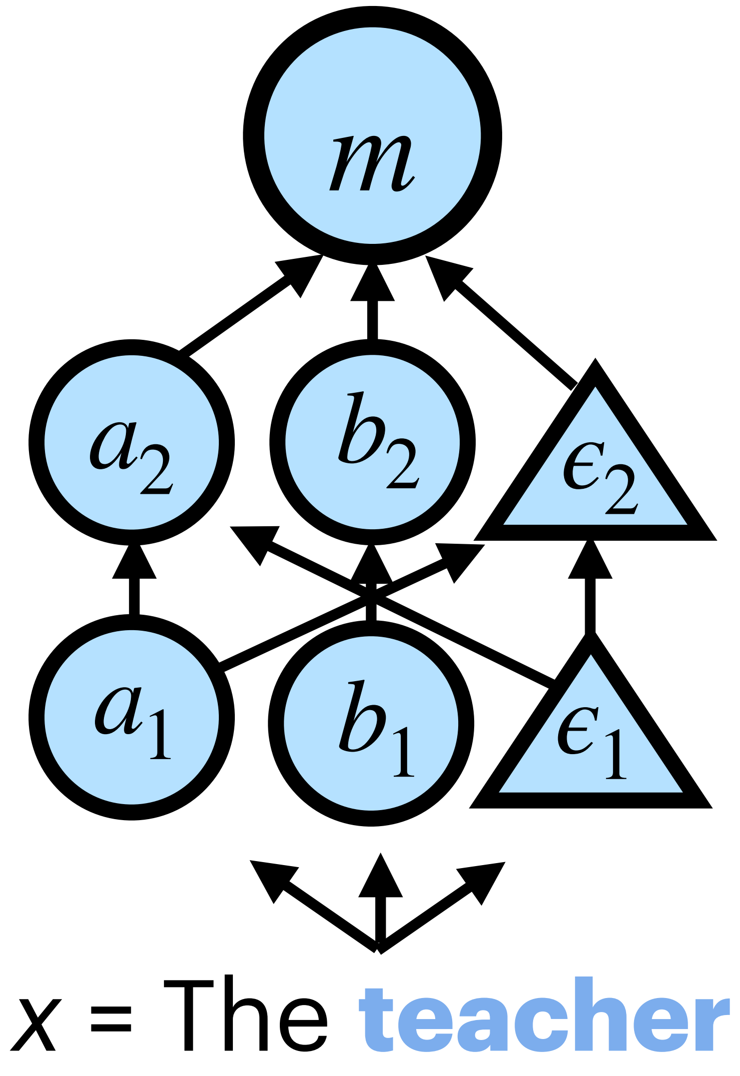
(We actually propose and use a more accurate approximation based on integrated gradients.)

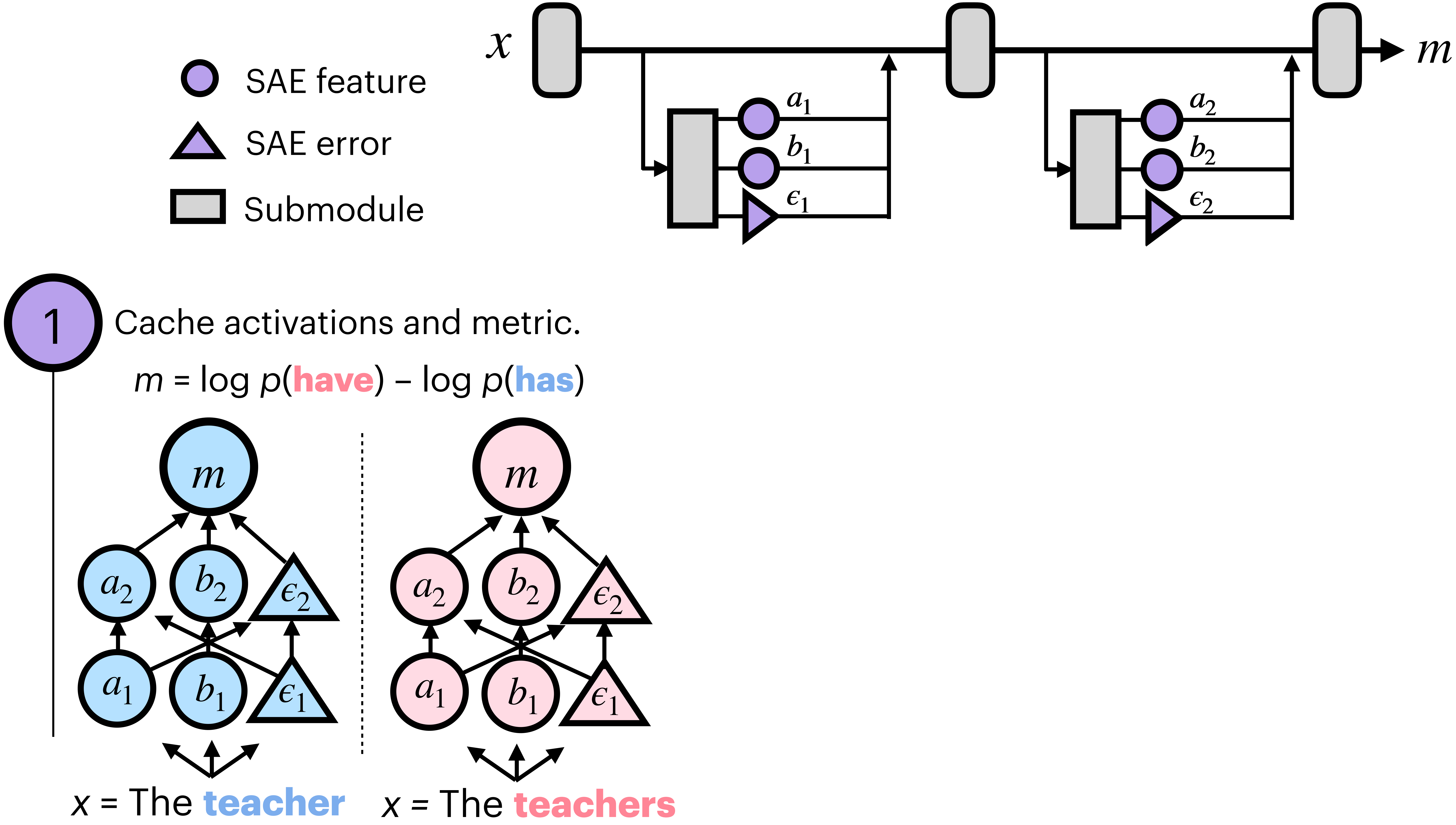


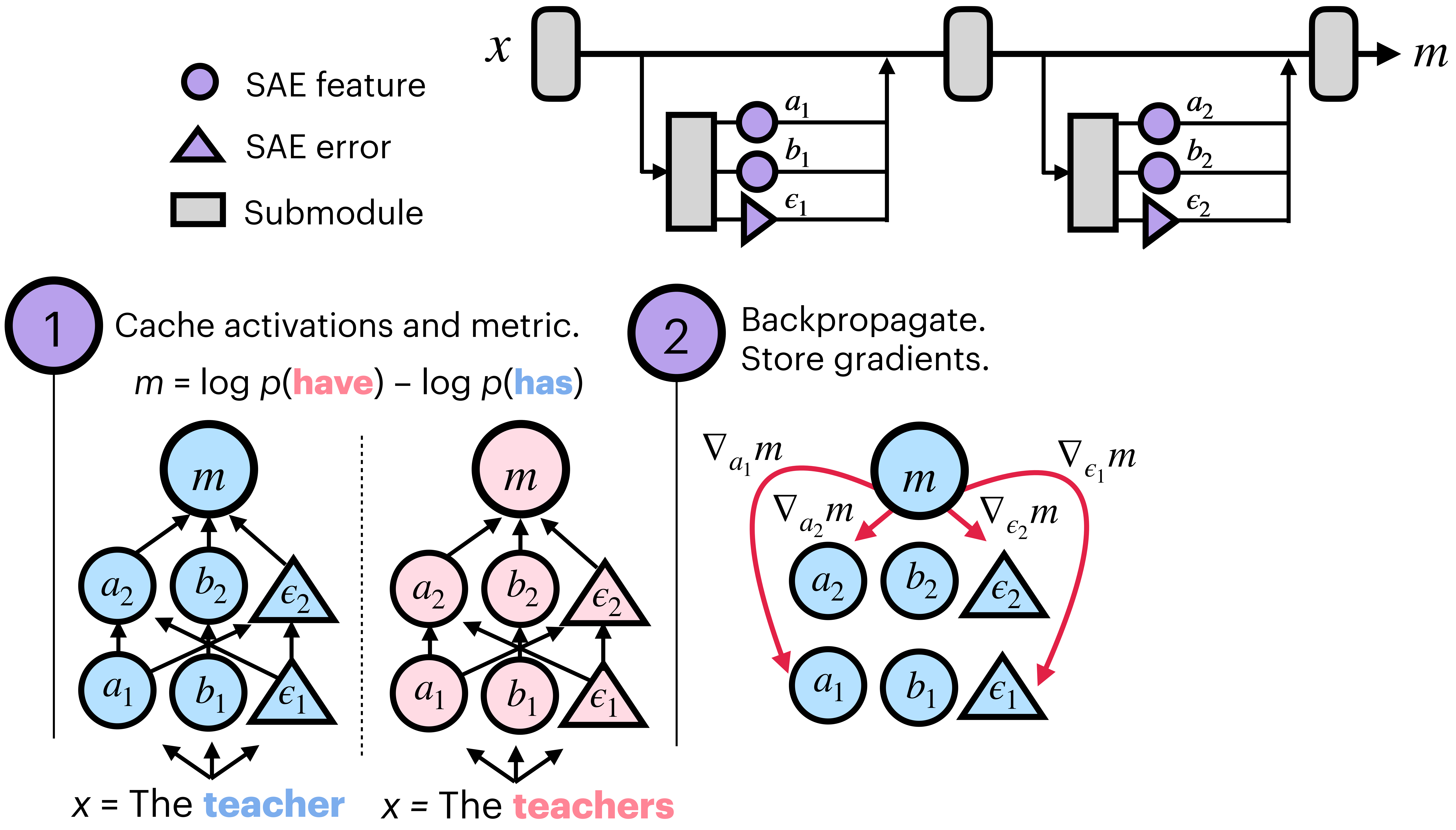
SAE feature
SAE error
Submodule
x
m
a_1
b_1
ϵ_1
a_2
b_2
ϵ_2

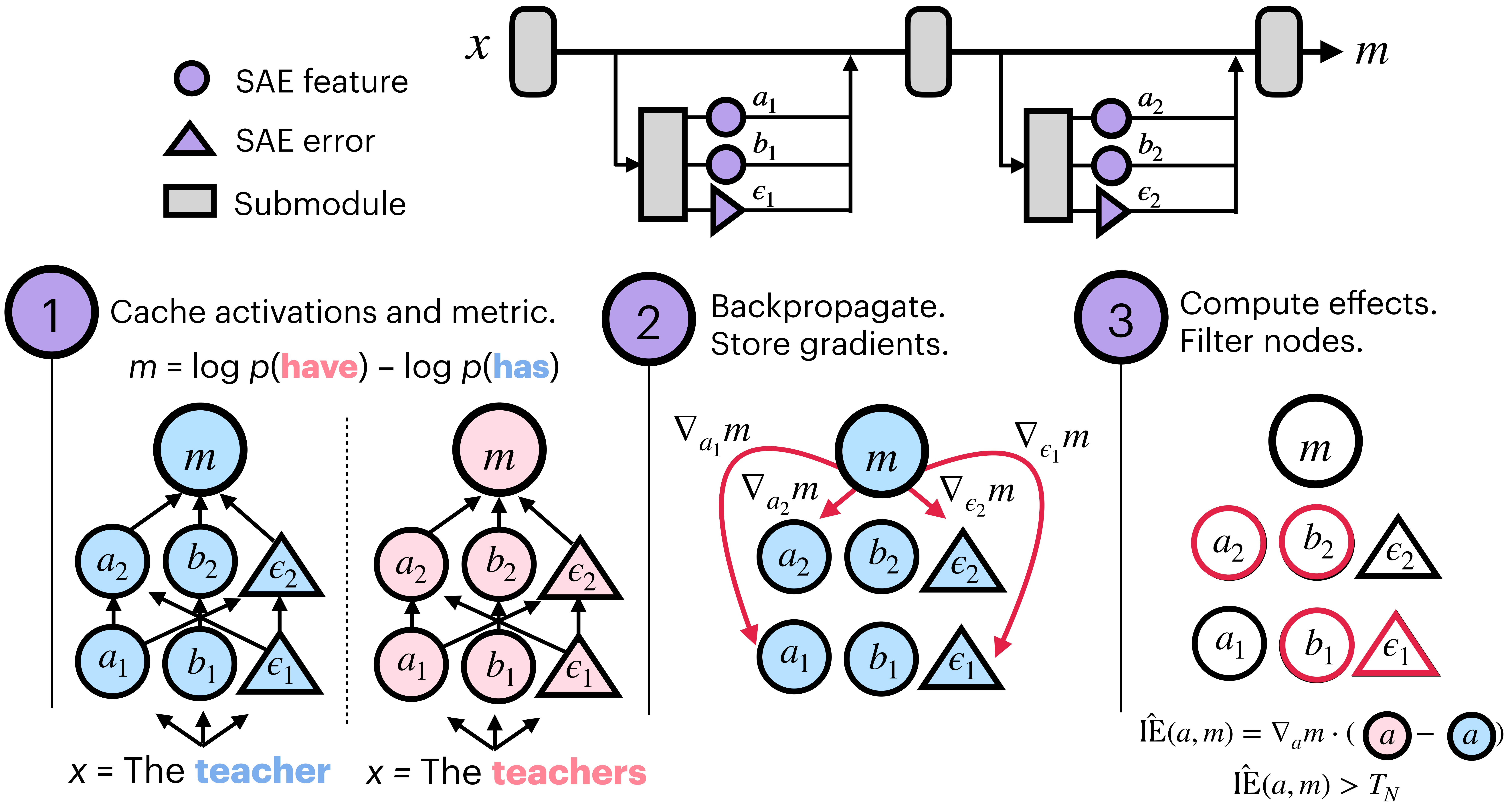


1 Cache activations and metric.

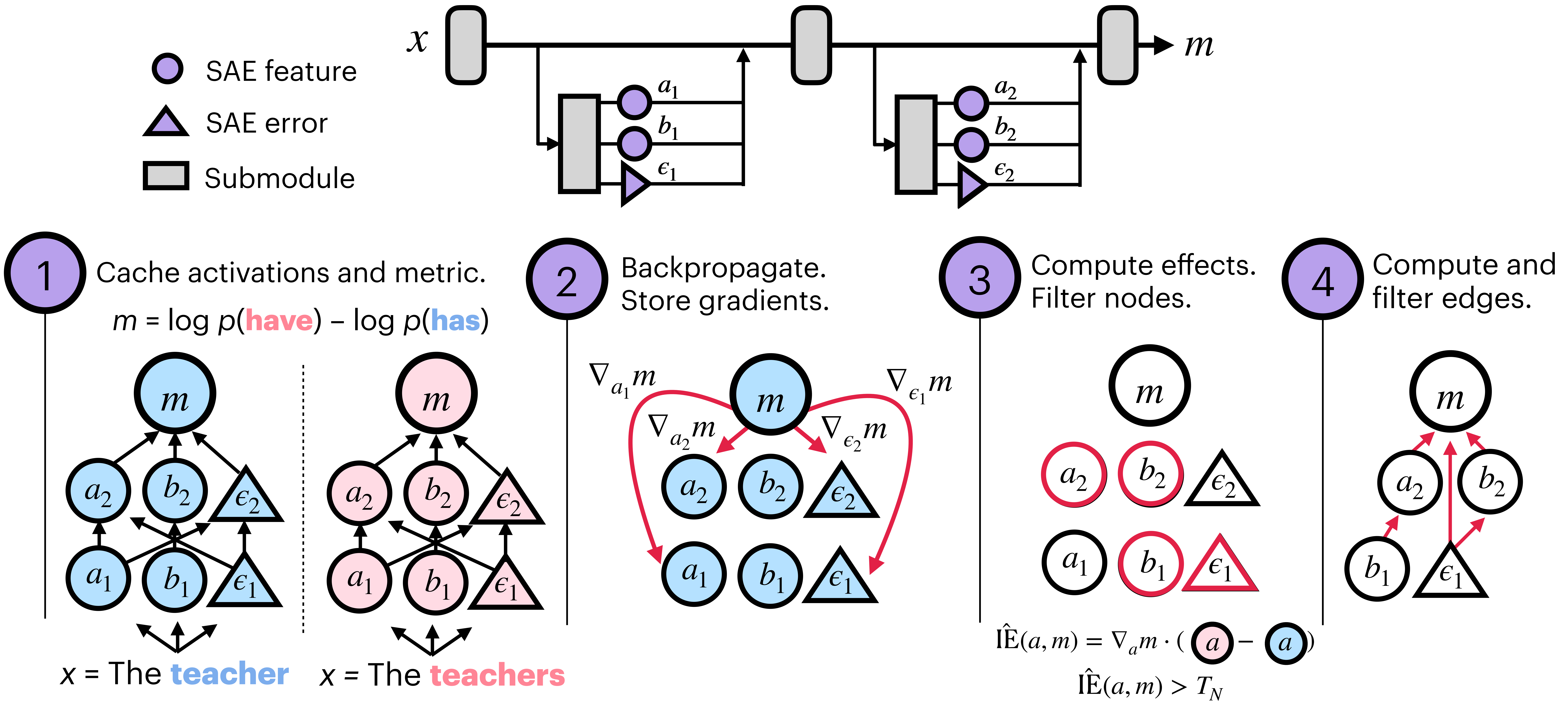



SAE feature
SAE error
Submodule
x
m
a1
b1
ϵ1
a2
b2
ϵ2
1
Cache activations and metric.
m = log p(have) – log p(has)
m
a2
b2
ϵ2
a1
b1
ϵ1
x = The teacher
x = The teachers


SAE feature
SAE error
Submodule
x
m
a1
b1
ϵ1
a2
b2
ϵ2
1
Cache activations and metric.
m = log p(have) − log p(has)
x = The teacher
x = The teachers
2
Backpropagate.
Store gradients.
∇a1 m
∇a2 m
∇ϵ2 m
∇ϵ1 m



SAE feature
SAE error
Submodule
x
m
a_1 b_1 ϵ_1
a_2 b_2 ϵ_2
1 Cache activations and metric.
m = log p(have) – log p(has)
x = The teacher
x = The teachers
2 Backpropagate.
Store gradients.
$\nabla_{a_1} m$
$\nabla_{a_2} m$
$\nabla_{\epsilon_2} m$
$\nabla_{\epsilon_1} m$
3 Compute effects.
Filter nodes.
$\hat{\mathrm{IE}}(a, m) = \nabla_a m \cdot (a - a)$
$\hat{\mathrm{IE}}(a, m) > T_N$


SAE feature
SAE error
Submodule
x
m
a_1 b_1 ϵ_1
a_2 b_2 ϵ_2
1 Cache activations and metric.
m = log p(have) – log p(has)
x = The teacher
x = The teachers
2 Backpropagate. Store gradients.
$\nabla_{a_1} m$
$\nabla_{a_2} m$
$\nabla_{\epsilon_2} m$
$\nabla_{\epsilon_1} m$
3 Compute effects. Filter nodes.
$\hat{\text{IE}}(a, m) = \nabla_a m \cdot (a - a)$
$\hat{\text{IE}}(a, m) > T_N$
4 Compute and filter edges.

Case Study

Subject–Verb Agreement

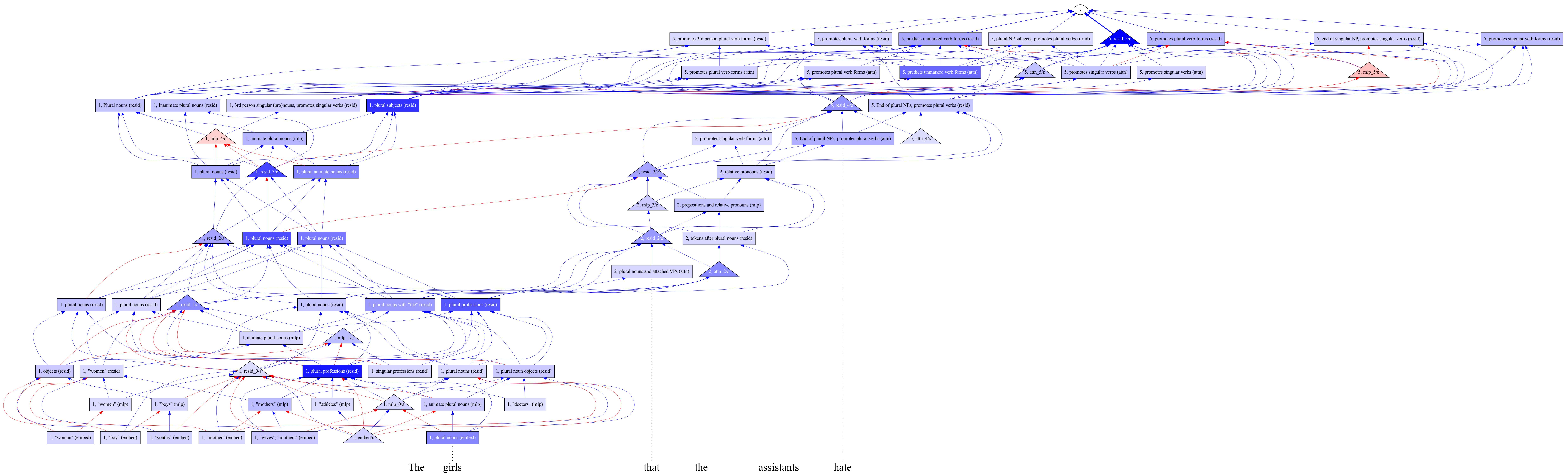
$$\mathrm{m} = p(\text{are}) - p(\text{is})$$

↑

x = The **manager** that the parents like

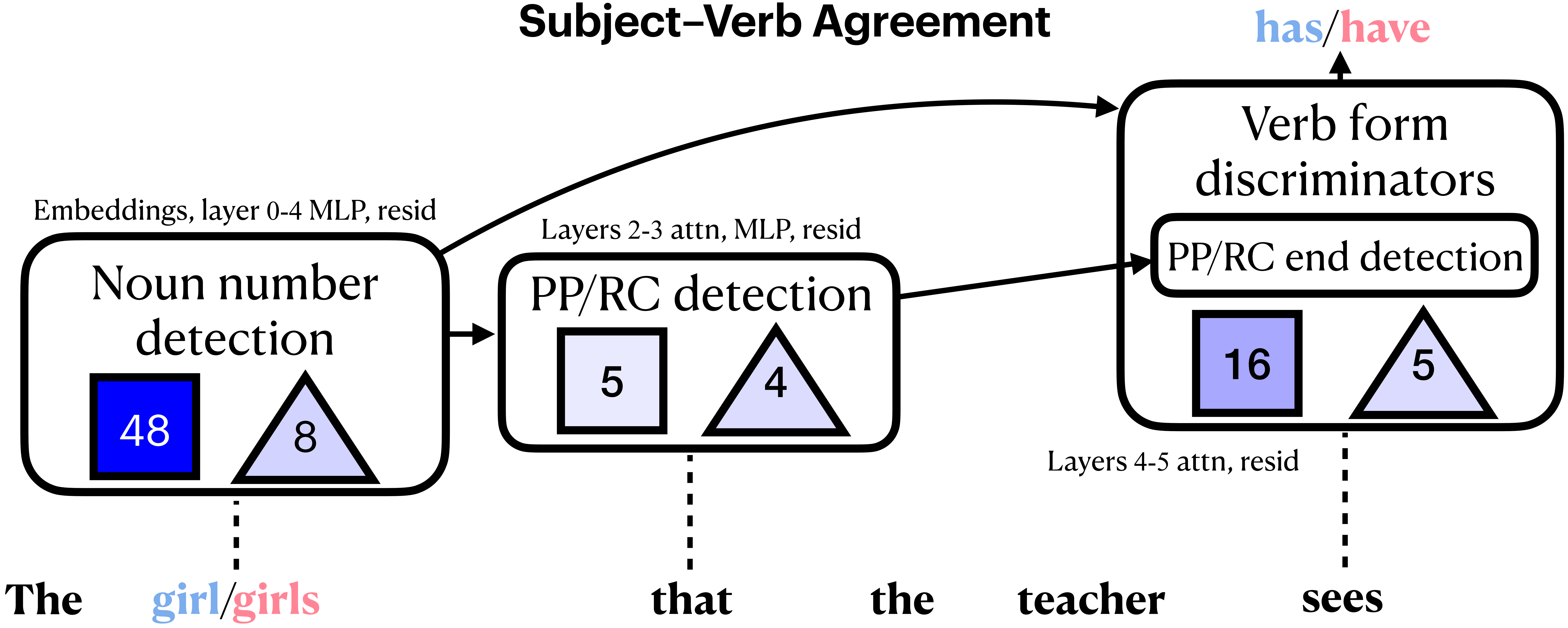
Case Study

Subject–Verb Agreement



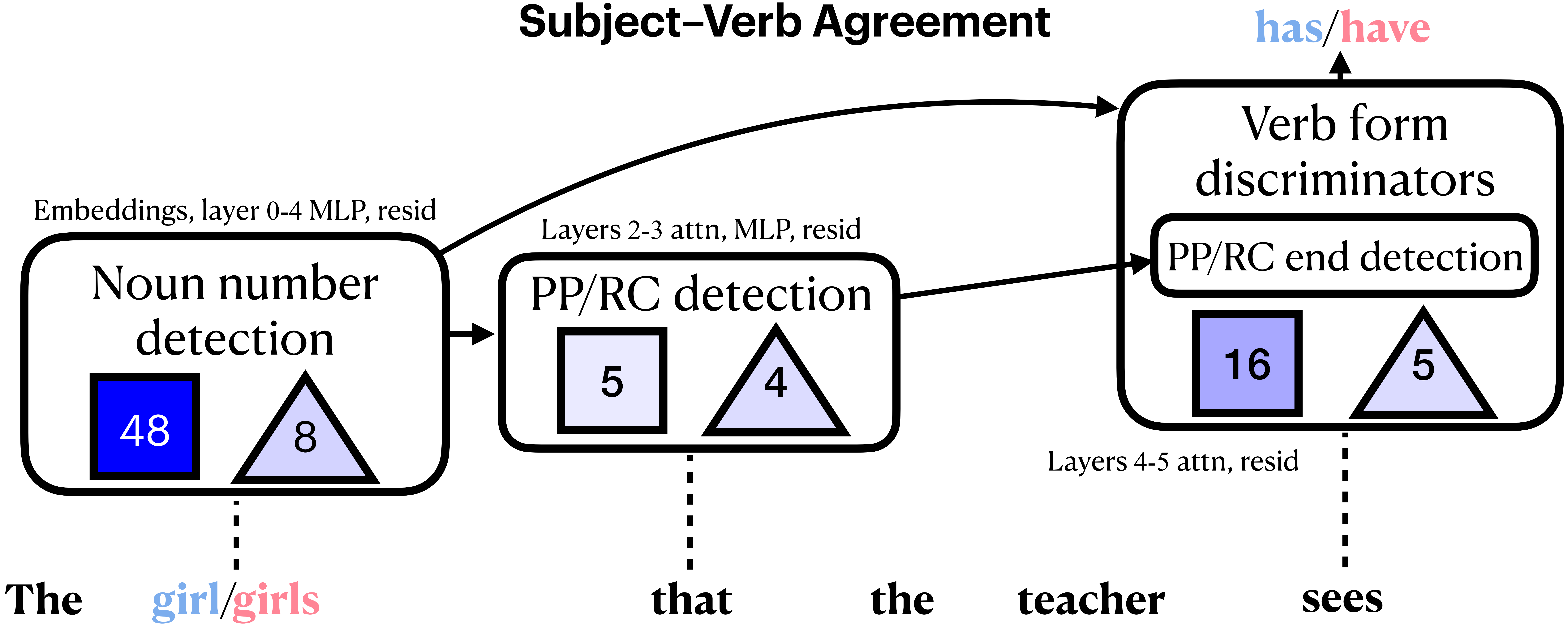
Case Study

Subject–Verb Agreement



Case Study

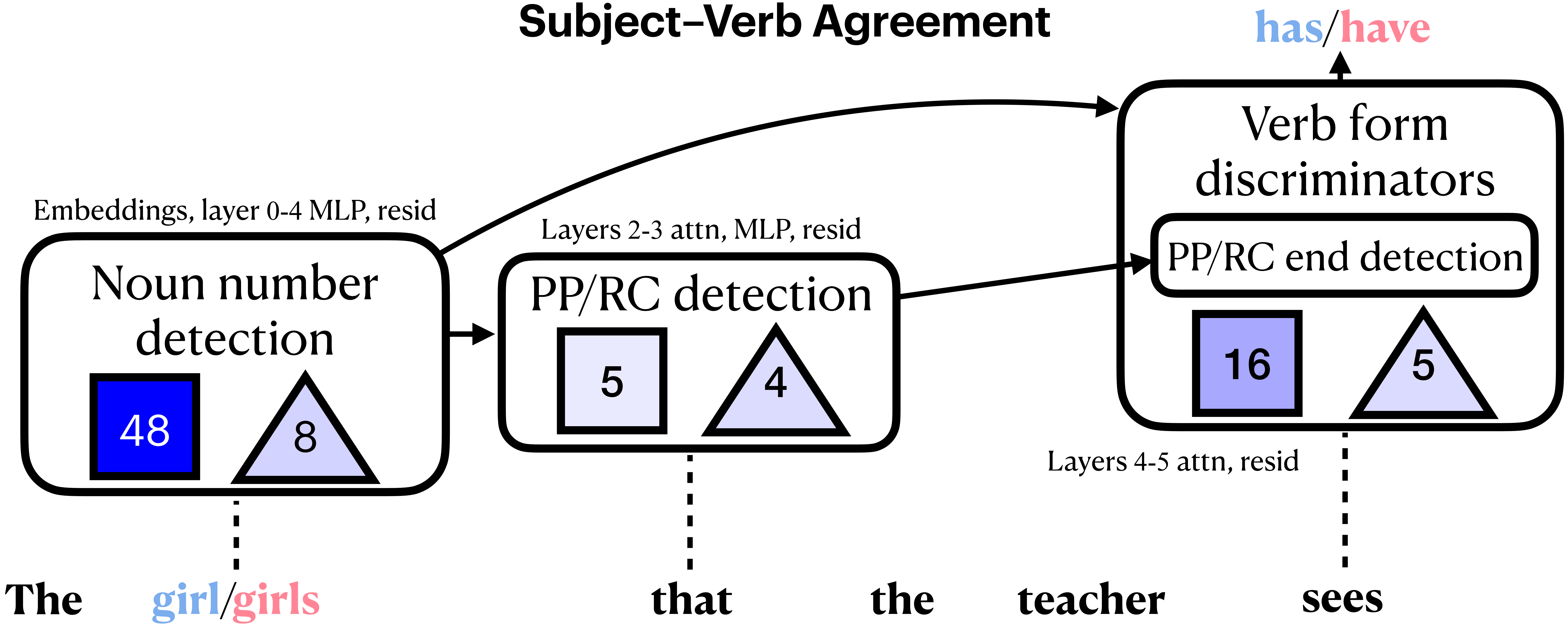
Subject–Verb Agreement



This corresponds to the human intuition!

Case Study

Subject–Verb Agreement



This corresponds to the human intuition!

But what about cases where it doesn't?

Classifying Ambiguous Data

Bias in Bios

Task: classify profession described in biography

Classifying Ambiguous Data

Bias in Bios

Task: classify profession described in biography

"He was previously an **assistant professor** at the University of Arizona..."	Professor 0
"She graduated in 2005 with honors, and has 11 years of experience as a **nurse practitioner**"	Nurse 1

Classifying Ambiguous Data

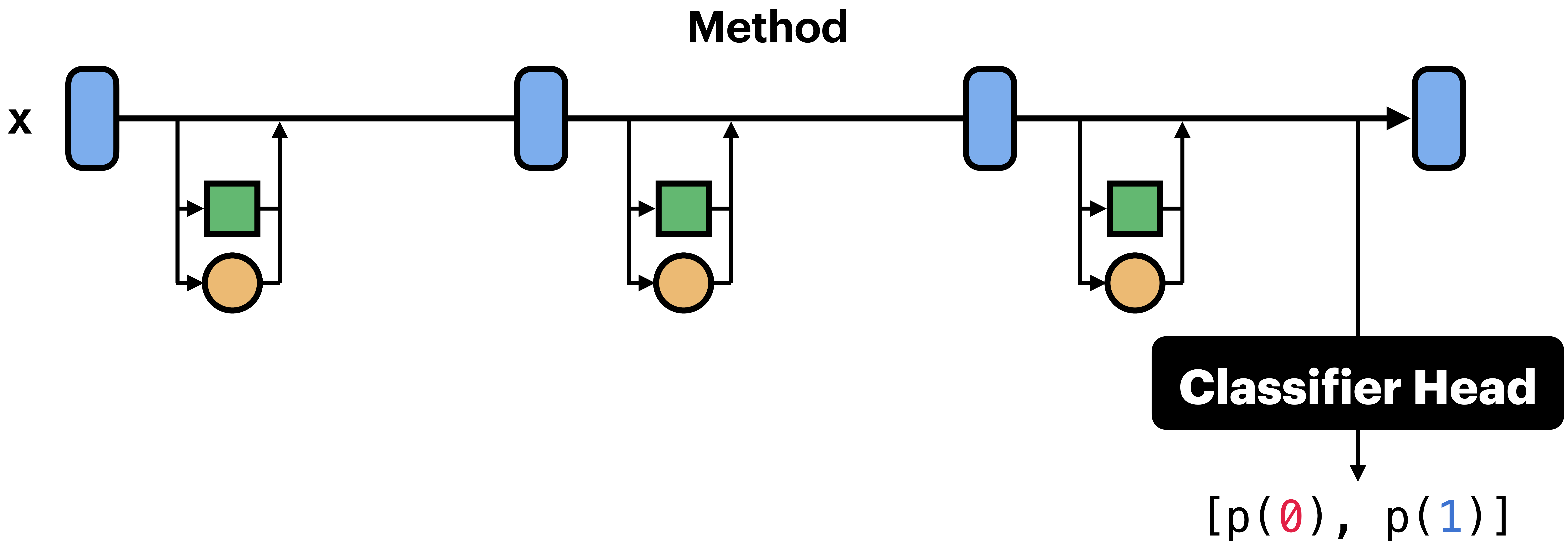
Bias in Bios

Task: classify profession described in biography

"**He** was previously an **assistant professor** at the University of Arizona…"	Man 0	Professor 0
"**She** graduated in 2005 with honors, and has 11 years of experience as a **nurse practitioner**"	Woman 1	Nurse 1

What if the target feature correlates perfectly with the spurious feature?

SHIFT



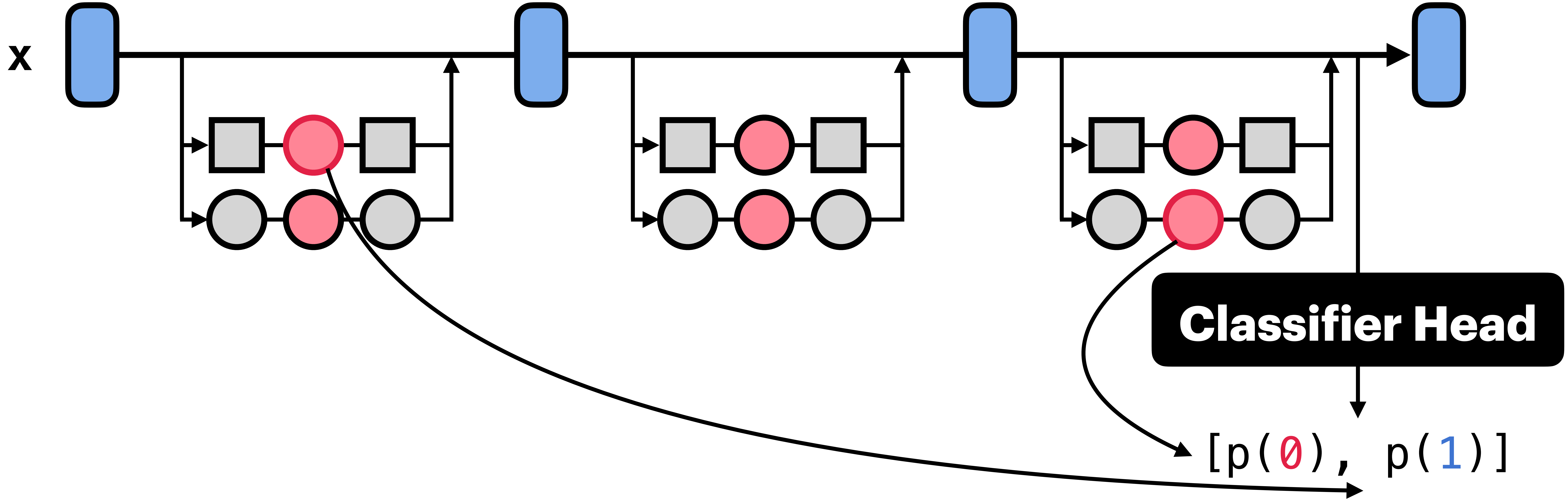
Task: classify profession

Acc.:

Profession : 63%
Gender: 87%

SHIFT

Method

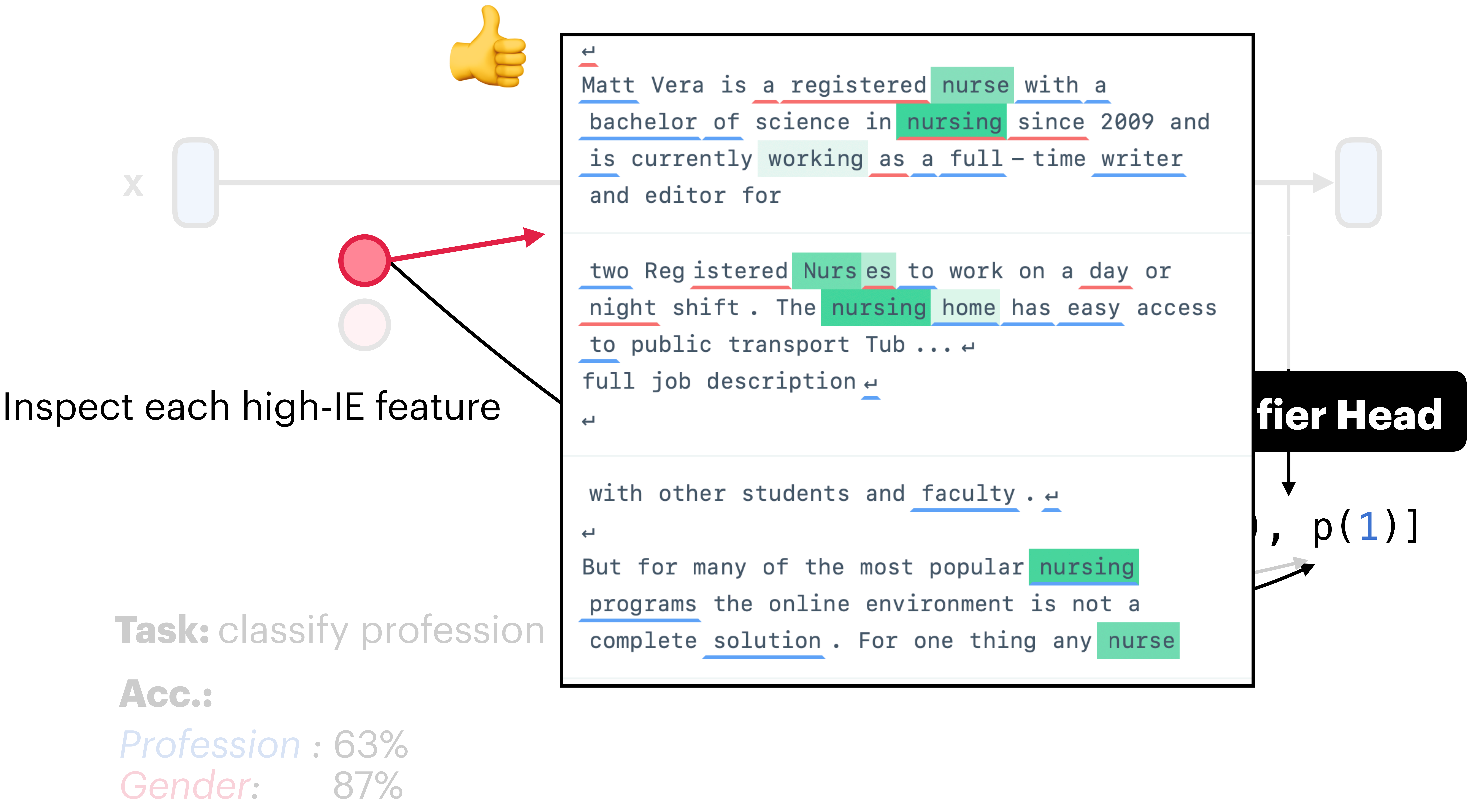


Task: classify profession

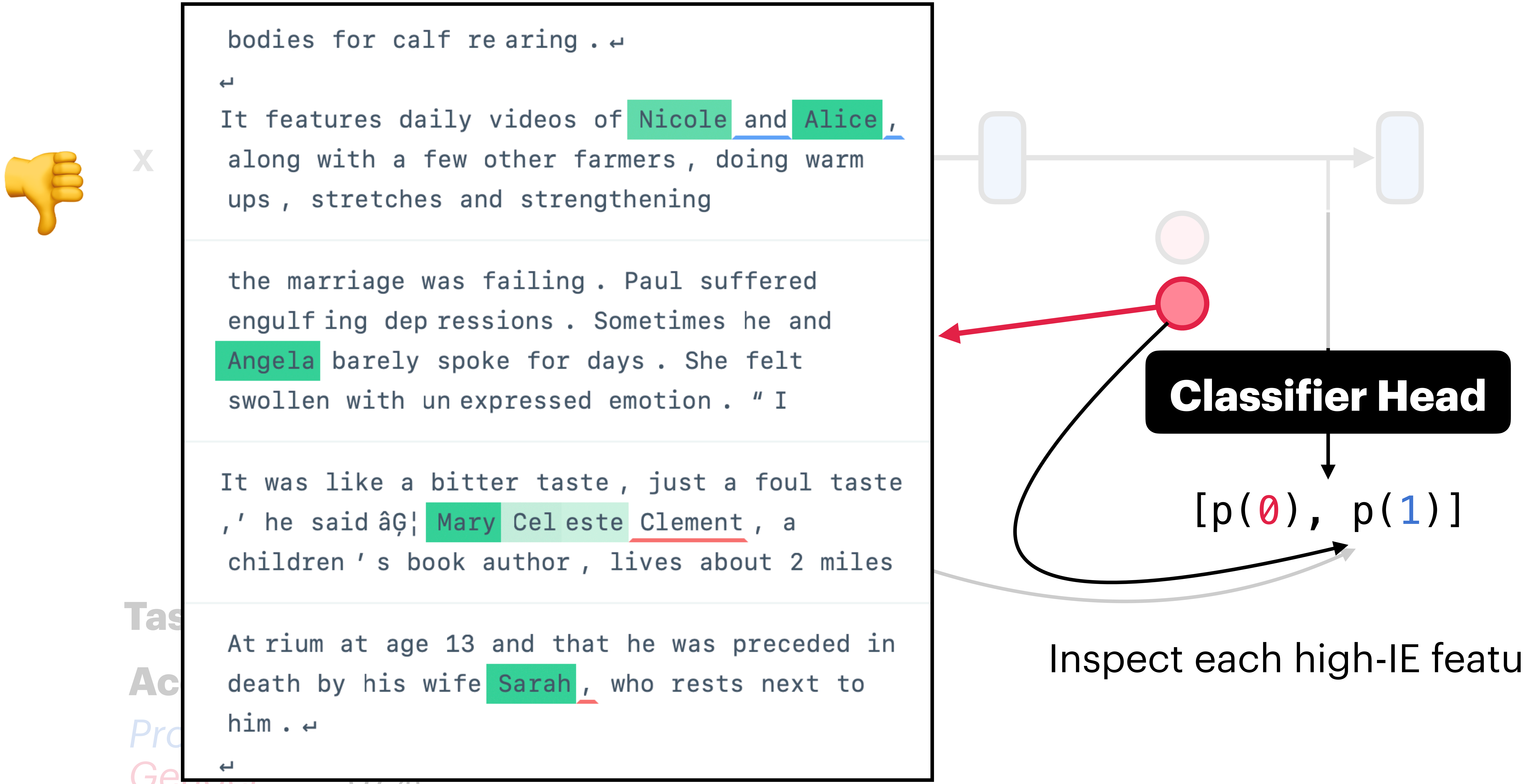
Acc.:

Profession : 63%
Gender: 87%

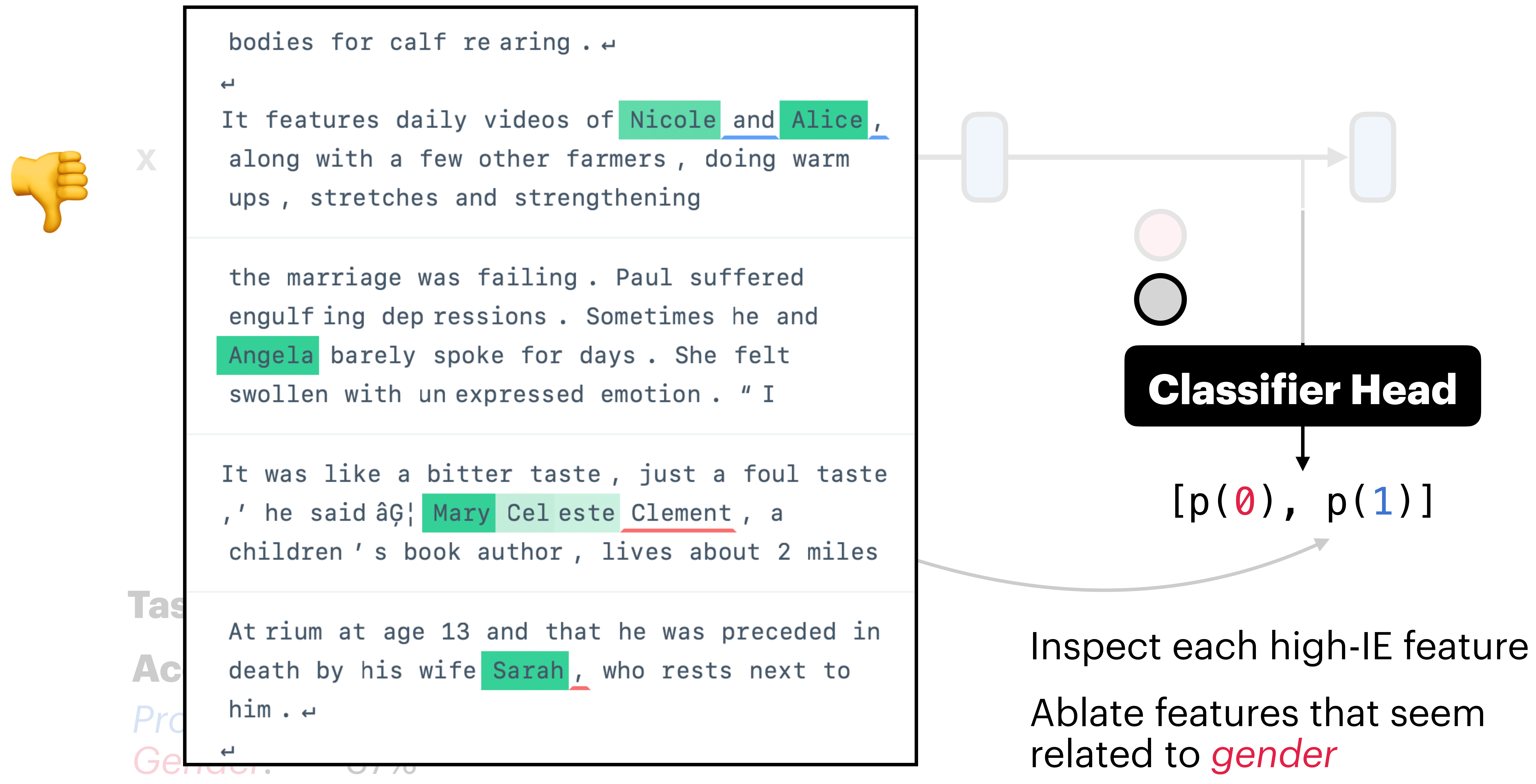
Look for features with high IE on classifier logits



Inspect each high-IE feature
x
fier Head
), p(1)]
Task: classify profession
Acc.:
Profession : 63%
Gender: 87%
↵
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 Reg istered Nurs es 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



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 . ↵
↵
Classifier Head
[p(0), p(1)]
Inspect each high-IE feature
Ablate features that seem related to *gender*

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
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**

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

	Pythia-70M			Gemma-2-2B		
Method	↑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!*

An Unsupervised Interpretability Pipeline

Intepretability typically requires us to have a behavior in mind.

Can we fully automate the behavior and circuit discovery process?

An Unsupervised Interpretability Pipeline

Intepretability typically requires us to have a behavior in mind.

Can we fully automate the behavior and circuit discovery process?

1. Given large text corpus $\{(x_i, y_i)\}$, collect activations of SAEs $\mathbf{v}(x_i, y_i)$
2. Cluster $\mathbf{v}$
3. Discover sparse feature circuits on clusters

An Unsupervised Interpretability Pipeline

Results

Cluster 382: Incrementing sequences

var input = [1, 2, 3, 4, 5, 6, 7, 8

Step 1. Download the latest CompsNY 3.49 Full
Step 2. Double click the Setup file and follow the prompts […]
Step 3. After the main install closes, click OK […]
Step 4

Cluster 475: “to” as infinitive object

At issue, whether the defendant should be allowed to

British Prime Min David Cameron says in televised remarks he would like Britain to

Reader bloggers are asked to

This yields not only interesting unanticipated behaviors…

An Unsupervised Interpretability Pipeline

Results

Cluster 382: Incrementing sequences

var input = [1, 2, 3, 4, 5, 6, 7, 8

Step 1. Download the latest CompsNY 3.49 Full
Step 2. Double click the Setup file and follow the prompts […]
Step 3. After the main install closes, click OK […]
Step 4

Example features involved:

Succession

- Chapter 1 Chapter 2 Chapter 3
- A, B, C
- I, II, III, IV

Narrow induction

- A3 … A → 3 or III or 4 …
- A7 … A → 7 or vii or 8 …

Cluster 475: “to” as infinitive object

At issue, whether the defendant should be allowed to

British Prime Min David Cameron says in televised remarks he would like Britain to

Reader bloggers are asked to

Example features involved:

Objects which can precede object complements

- Direct the user to
- It’s up to you to

Other words which precede infinitive objects

- According to
- This infection leads to

This yields not only interesting unanticipated behaviors…

…But also interesting unanticipated features!

Takeaways

1. Sparse feature circuits allow us to derive **human-interpretable** and **editable** causal graphs from LMs.

2. They allow us to surgically improve model generalization *without additional data*.

3. They allow us to automatically discover **unanticipated** model behaviors and mechanisms.

Thank you!

🌐 Project Website

Check out our poster:

Today 3pm – 5:30pm
Poster #495

Open Philanthropy
NSF
ISRAEL SCIENCE FOUNDATION
Fondation Azrieli Foundation
Mortimer B. Zuckerman STEM Leadership Program






Sparse Feature Circuits

Discovering and Editing Interpretable Causal Graphs in Language Models

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, **Aaron Mueller**

2025 International Conference on Machine Learning

26 April 2025