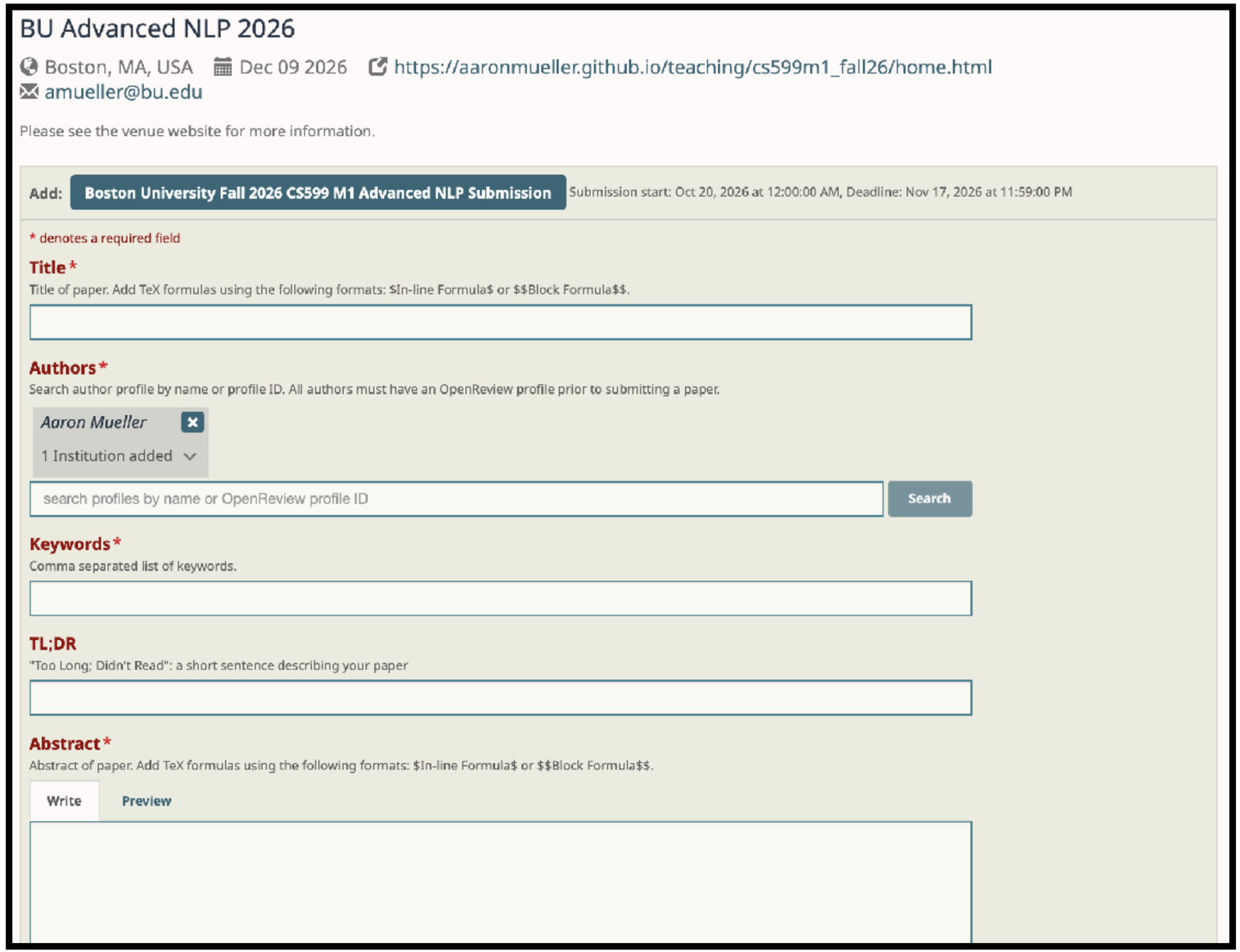
Interpretability II

Aaron Mueller
CS599 B1: Advanced Natural Language Processing
Nov. 10, 2026

Midway Report Submission



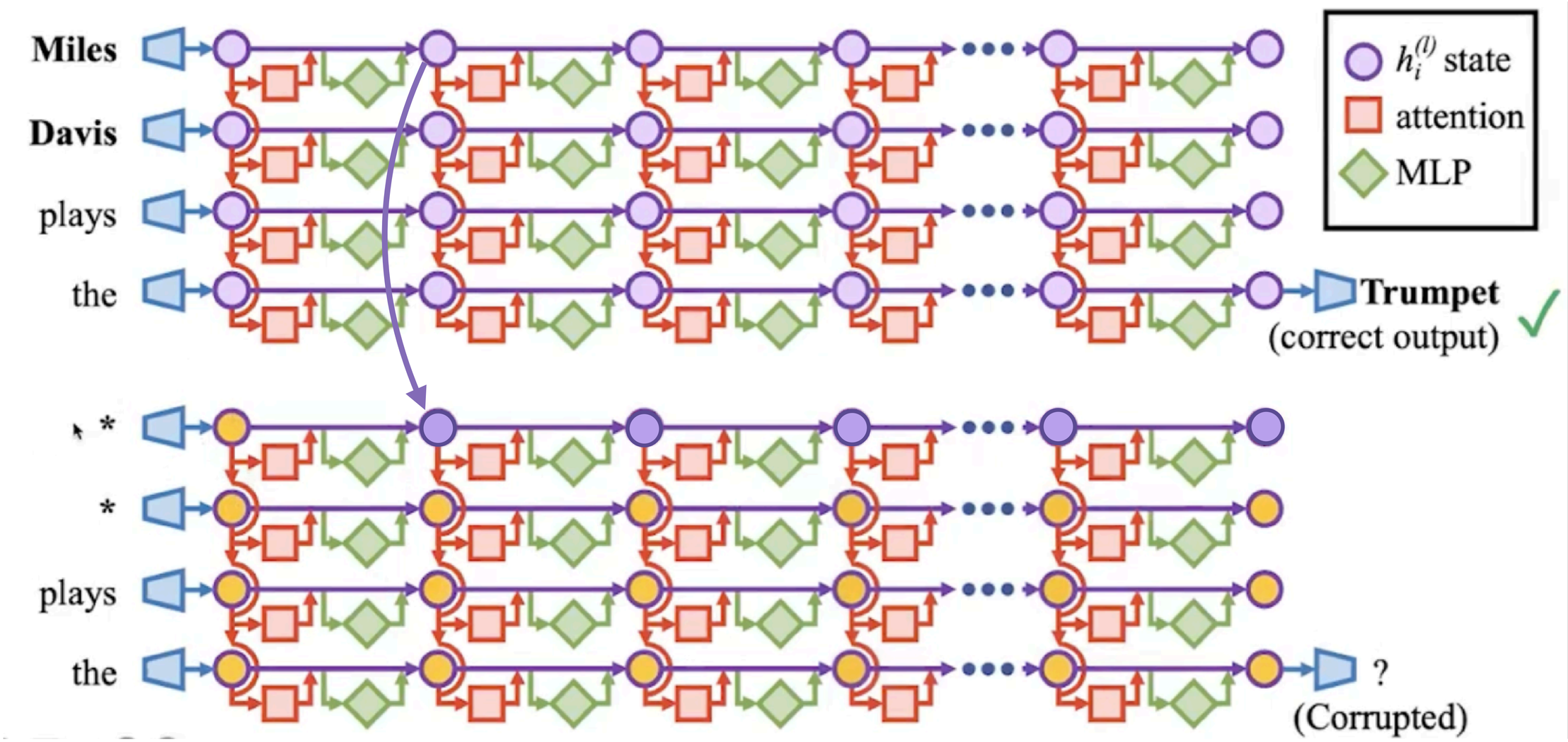
https://openreview.net/group?id=bu.edu/Boston_University/Fall_2026/CS599_M1_Advanced_NLP

Due **Thursday, Nov. 12**

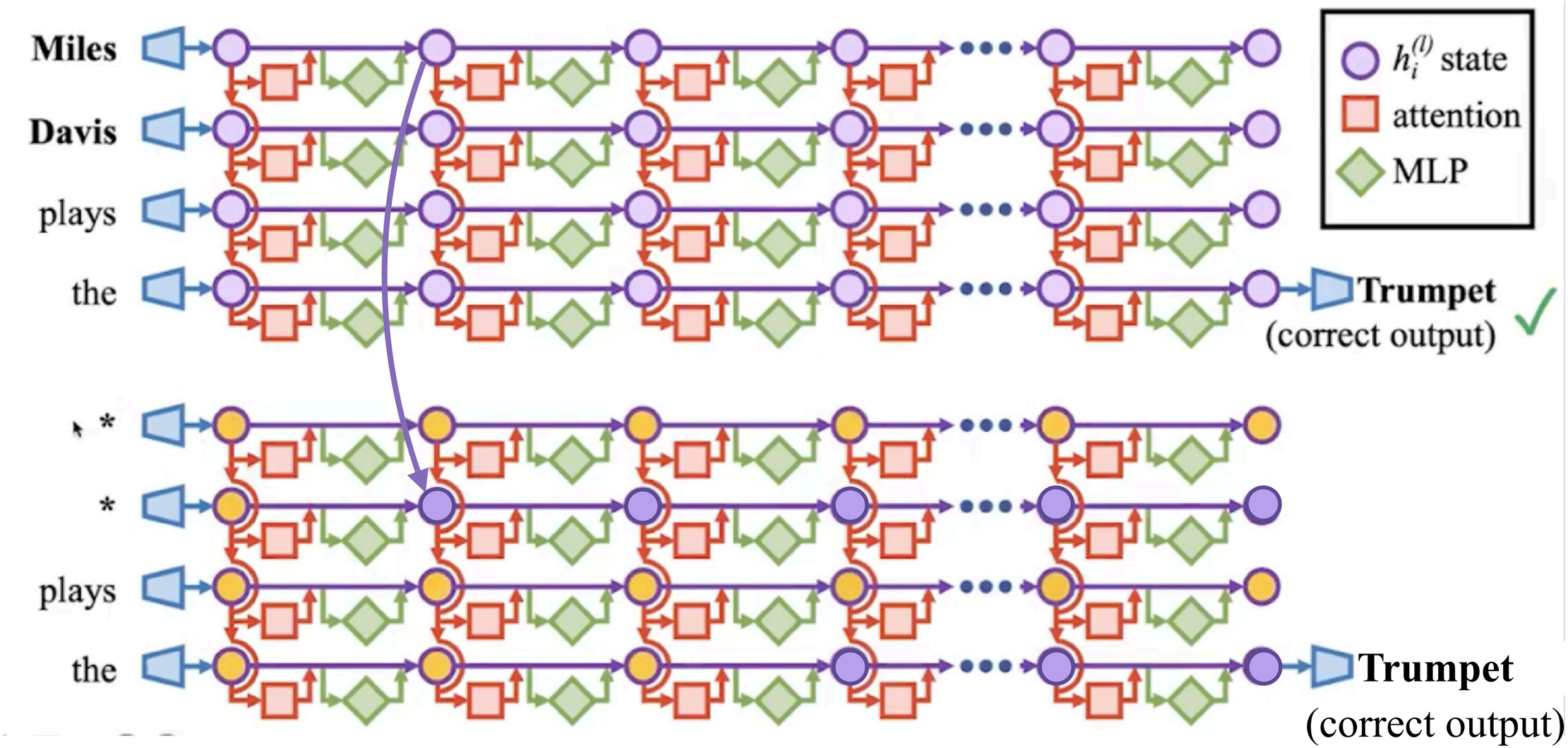
Make sure your PDFs are anonymized!

Also, be sure to make your OpenReview accounts soon!

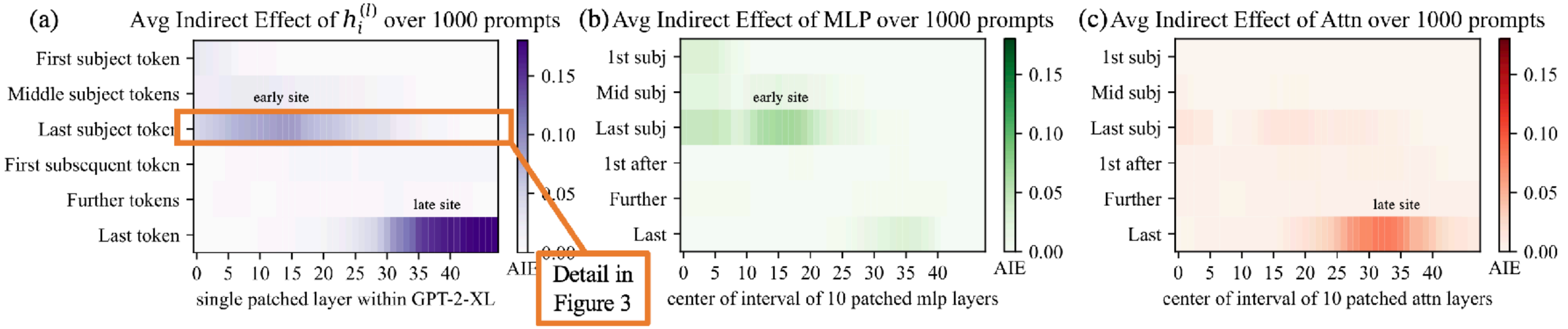
Locating Knowledge in LMs



Locating Knowledge in LMs



Causal Tracing



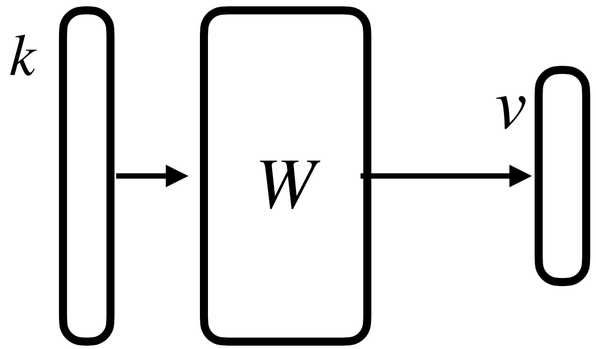
- Knowledge is encoded quite consistently across facts:
  - Two sites
  - Last subject token
  - Mid-layer MLP and late-layer attention modules do a lot of the work

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. $Wk_i \approx v_i$

Editing Knowledge in a Neural Layer

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

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

- 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:

$$W_1 = \arg\min_W \|V_1 - WK_1\|^2$$

- 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

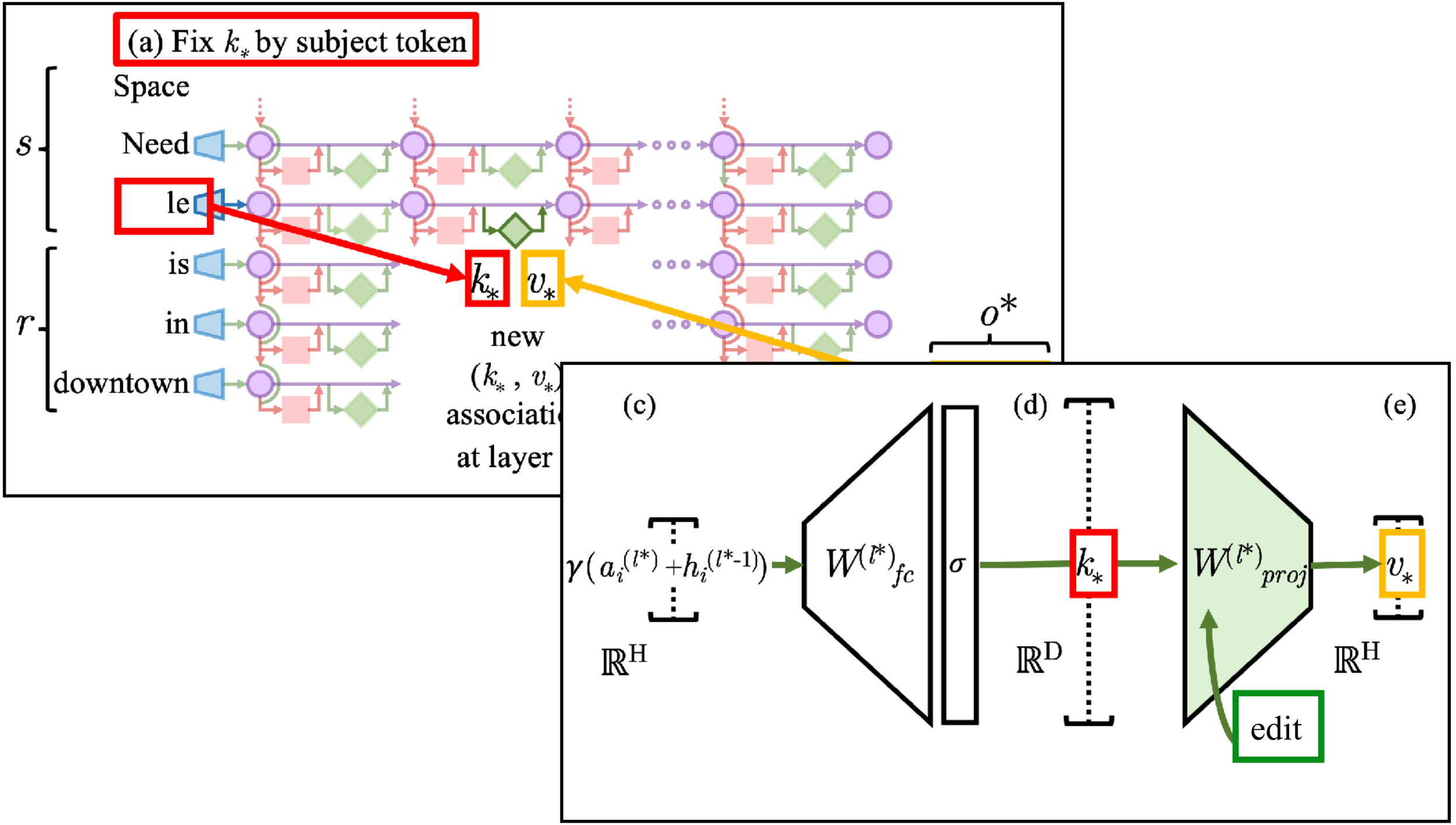
desired output

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

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

old weights

Gram matrix of new key


(a) Fix k_* by subject token
Space
s
Need
le
r
is
in
downtown
k_*
v_*
new
(k_*, v_*)
associati
at layer
o^*
(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

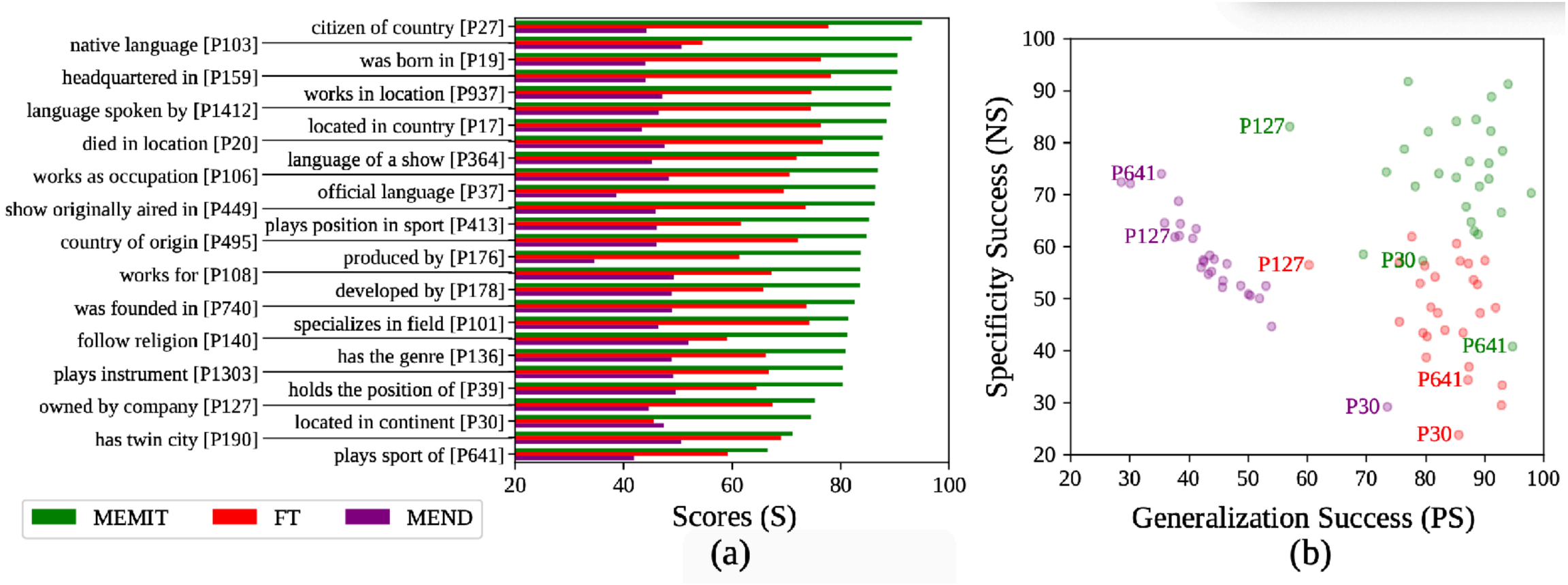
<u>Added fact: The Eiffel Tower is in Rome.</u>

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)



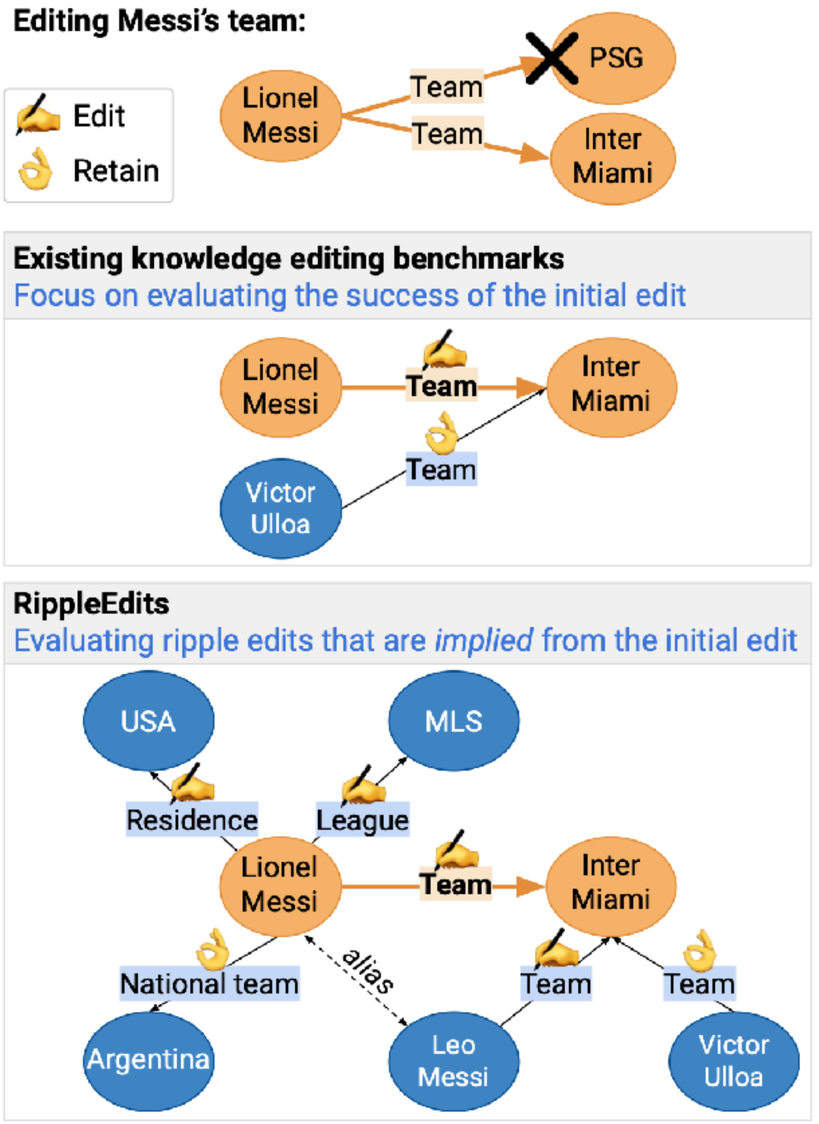
Knowledge editing (MEMIT) seems to work better than fine-tuning (FT)!

There is a trade-off between generalization and specificity.

Known Challenges in Model Editing

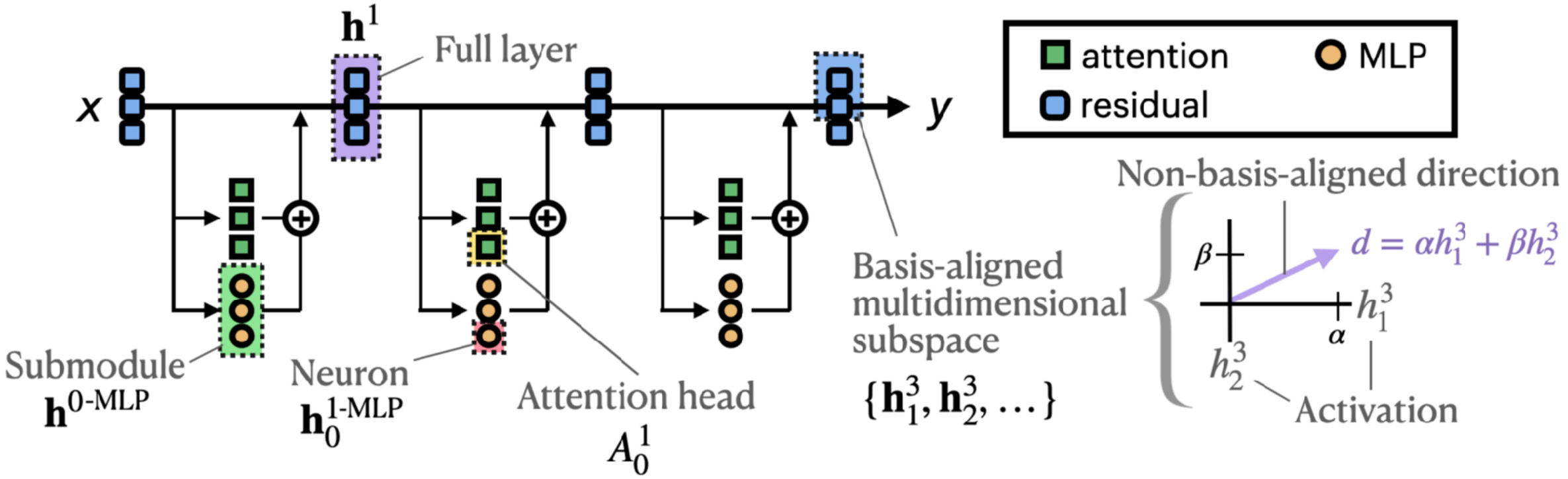
- It's been found that edits often don't generalize to logically related facts.

- Some conceptualize model editing as operating over token association probabilities, rather than latent knowledge *per se* **[Niu et al., 2024; Hase et al., 2023]**.
    - (This is perhaps too pessimistic, but many known failure modes could be explained by this perspective.)



What are the right mediators?

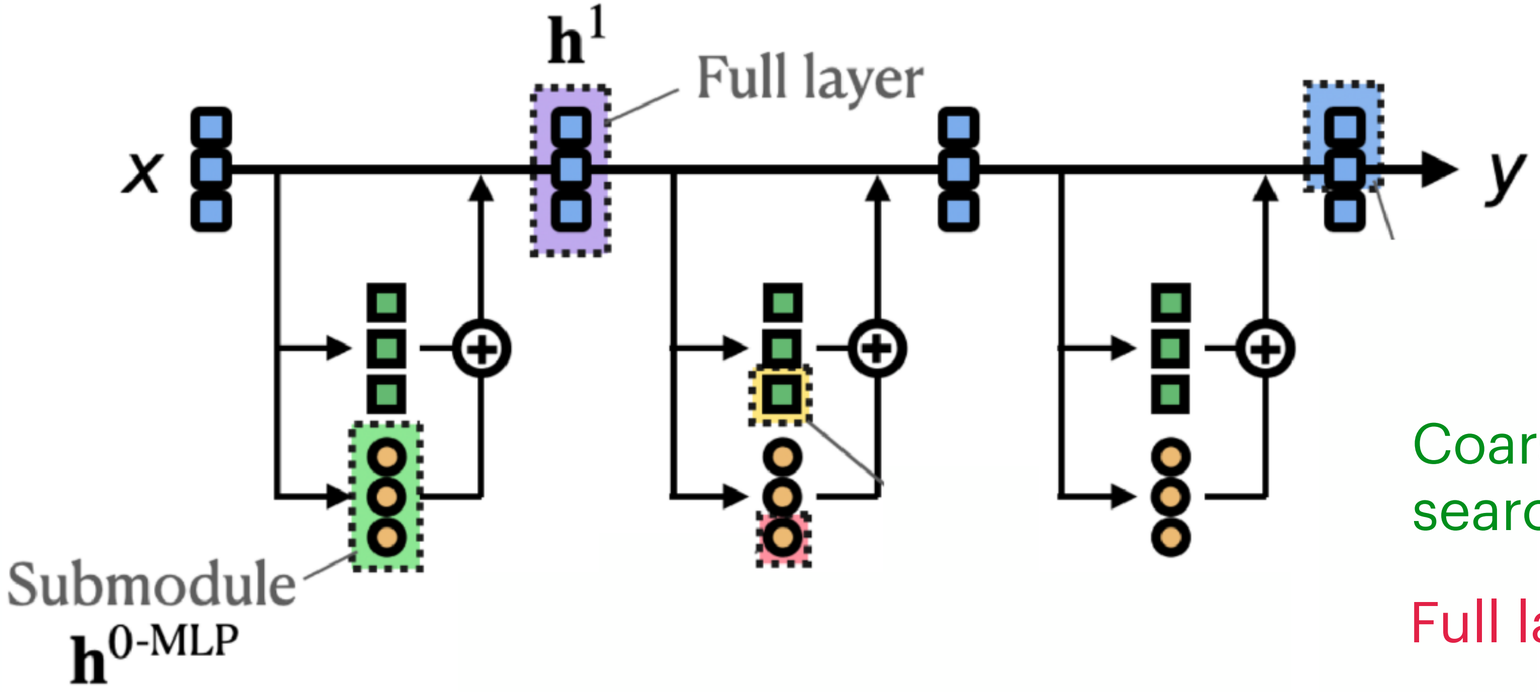
How do we know where to look for a given concept, like gender, or parts of speech, or other concepts?



[Mueller et al., 2026]

What are the right mediators?

What if we try intervening on a full layer (i.e., all neurons in a given layer)?





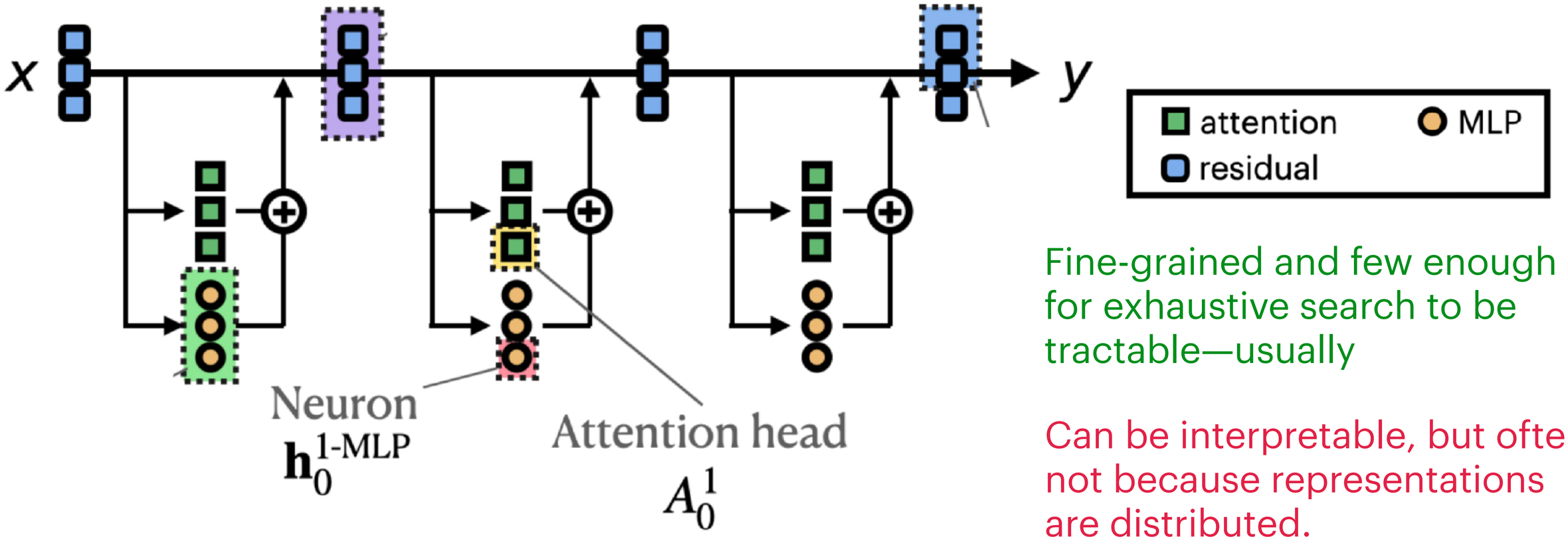
Coarse-grained, so exhaustive searches are fast!

Full layers are **causal bottlenecks**

Not super interpretable

What are the right mediators?

What about single neurons, or single attention heads?

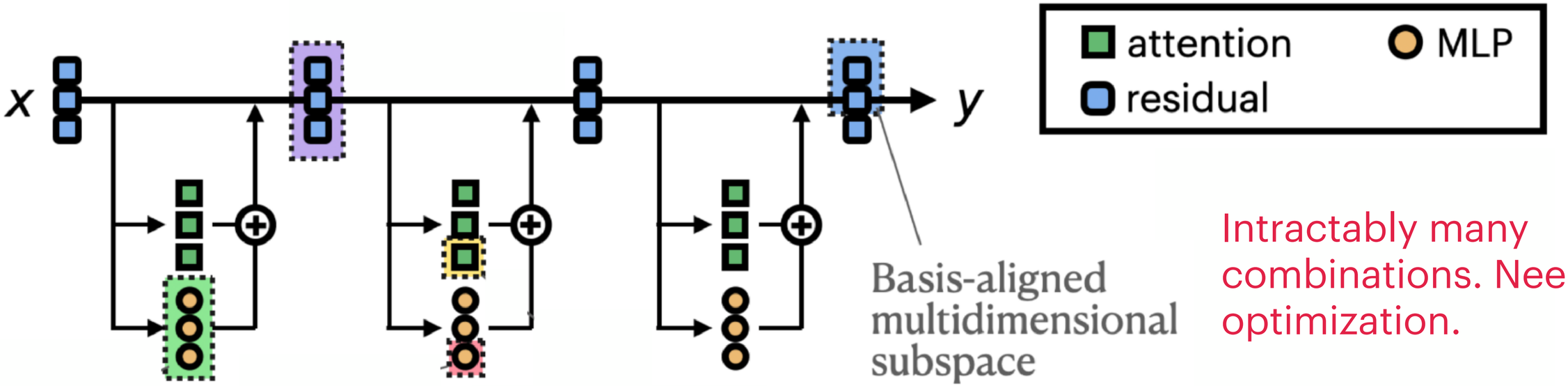


Fine-grained and few enough for exhaustive search to be tractable—usually

Can be interpretable, but often not because representations are distributed.

What are the right mediators?

What about *groups of* neurons or attention heads?

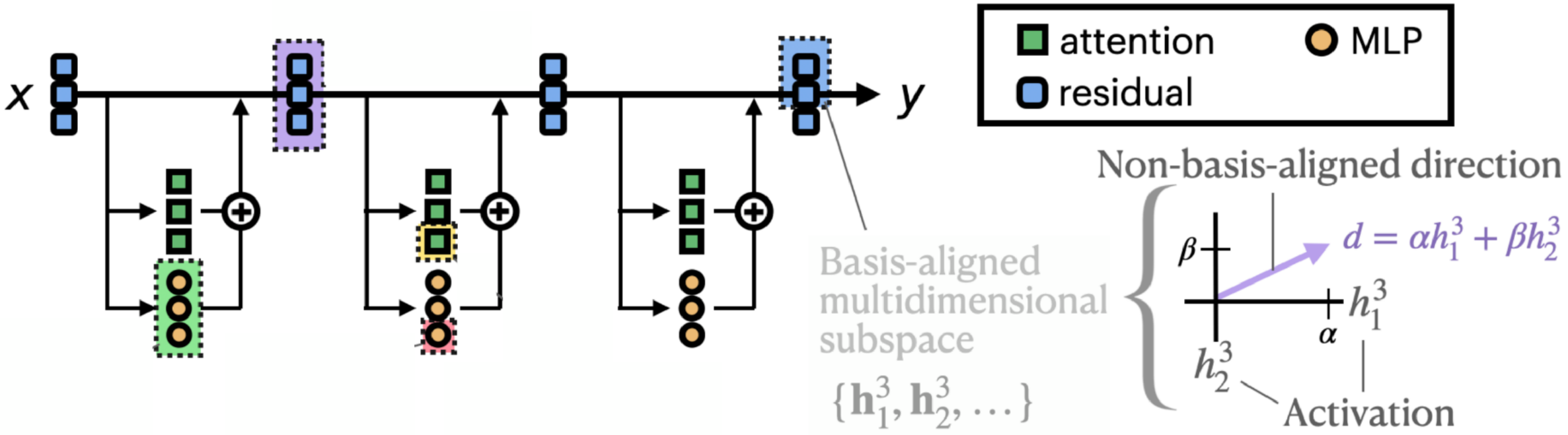


Intractably many combinations. Need optimization.

Generally not super interpretable.

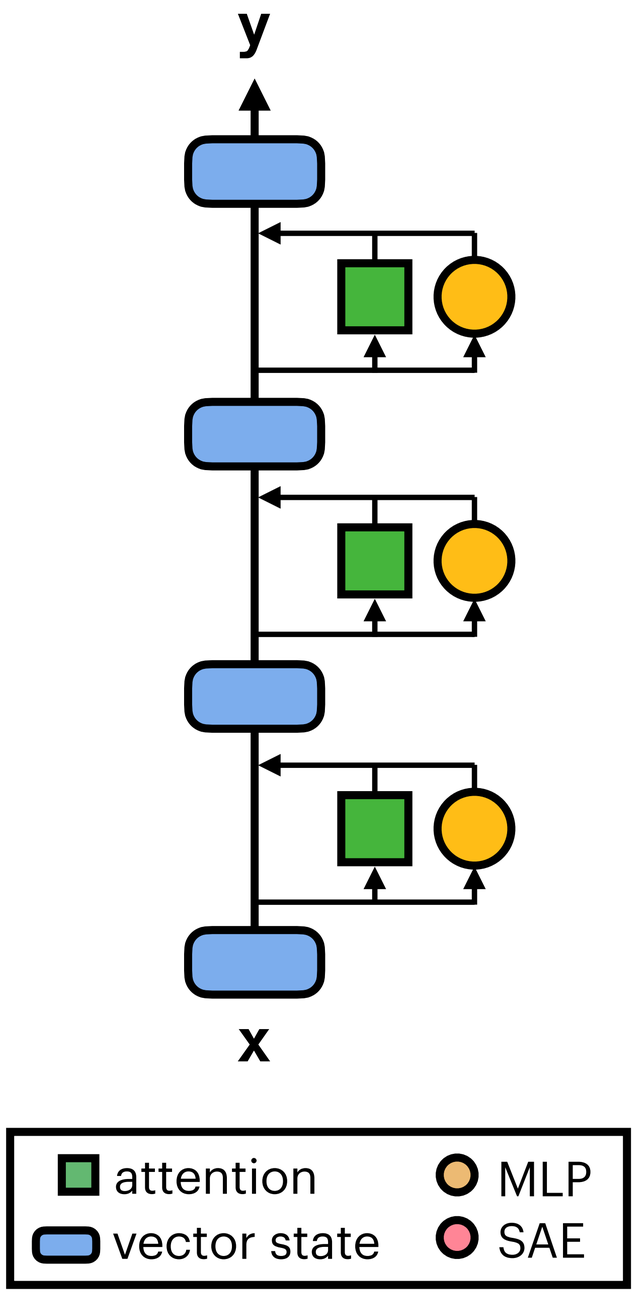
What are the right mediators?

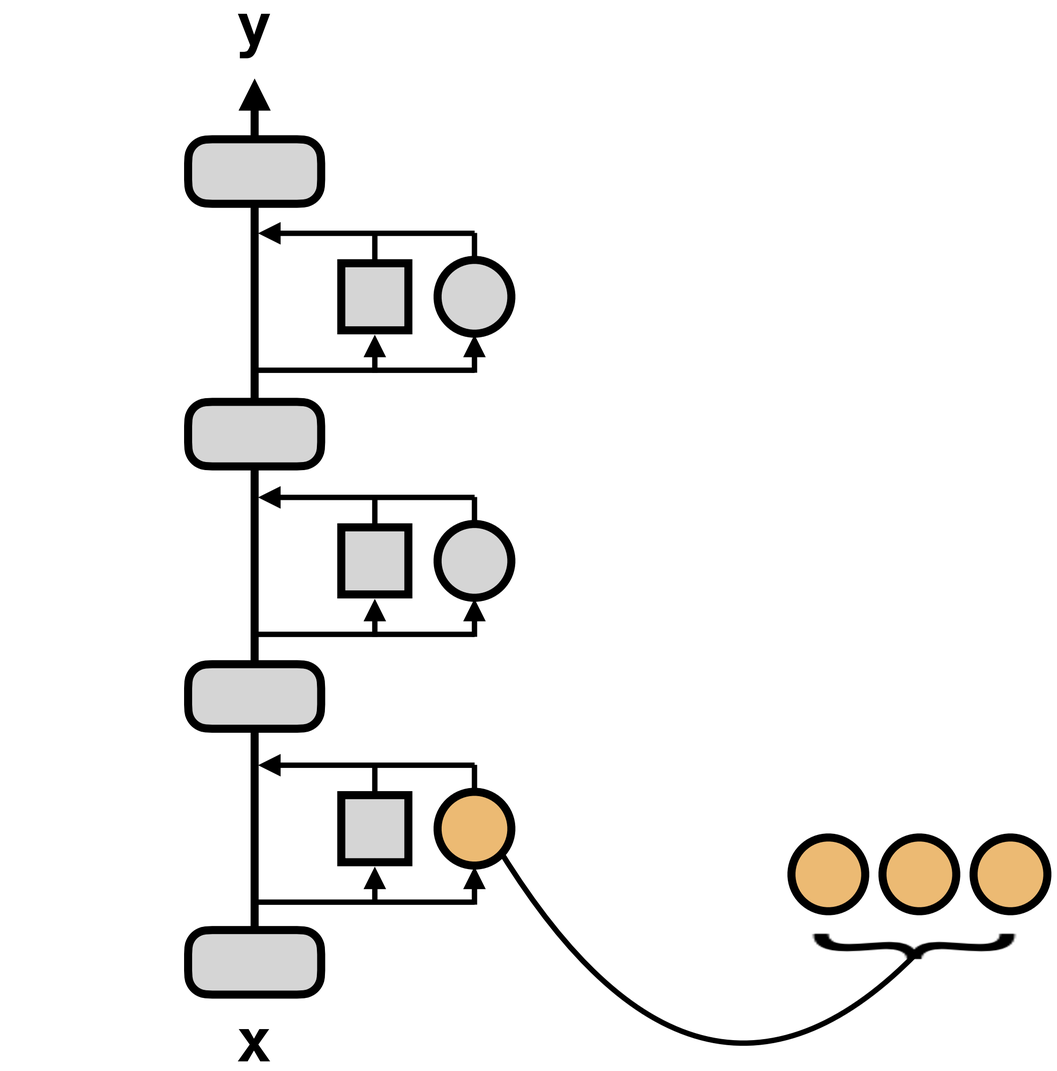
What about directions (weighted combinations of neurons)?



Infinitely many combinations.
Need optimization.

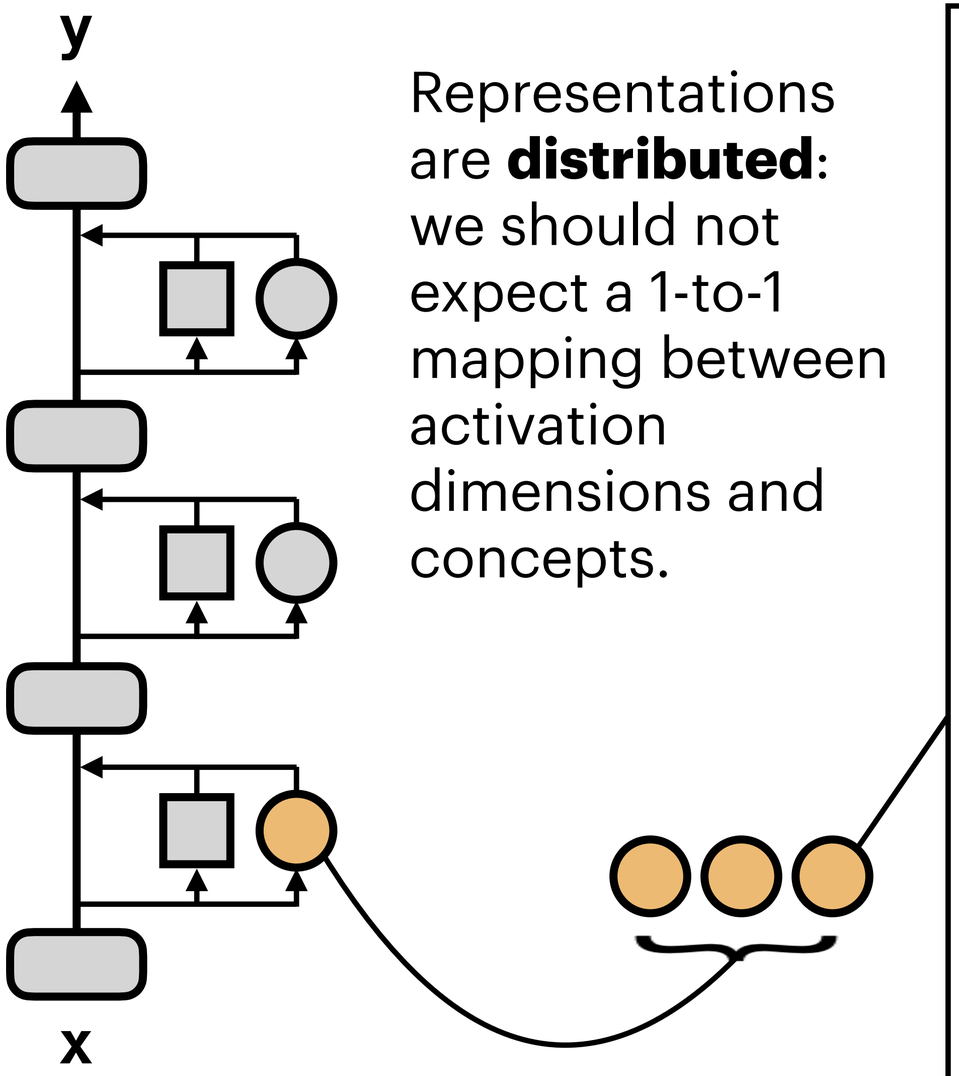
Fine-grained, and can be interpretable!


y
x
attention
MLP
vector state
SAE



Do neurons encode concepts?

[Bricken et al., 2023]



Representations are **distributed**: we should not expect a 1-to-1 mapping between activation dimensions and concepts.

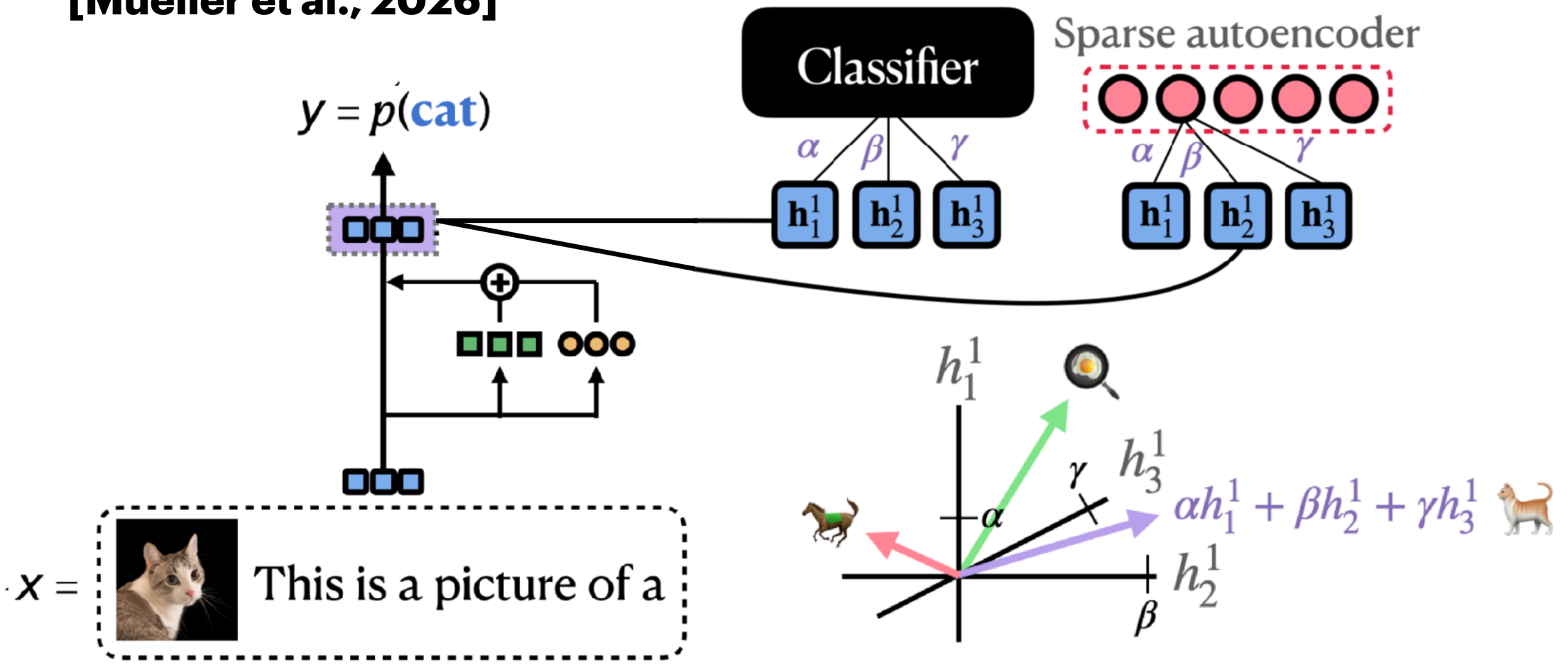
TOP ACTIVATIONS
TRAIN TOKEN MAX ACT = 5.406

particles random walk in one dimension until they strike
, D**).](zookeys-839-
\xe2\x86\x90
use [Backspace ←] and
me/haxe-three.js⏎
each from Saxe-Meiningen,
; The fourth box-plot corresponds to the
rongramteepuck.com/public
UL, 0xe02bd389UL,
IN predicate in Chapter 9 that a list of
ink-jet printhead (not seen in
: the text insideexample_textis a
print(new_zoo)⏎('monkey
0xe147ae10},␣␣␣/*
chess of Saxe-Coburg and
⏎<EOT>Photos from one guy's travels around
substituting the B12345678 with your actual Kindle
comma select records from one matching the left expression
a search box like flipkart website.
iruvananthapuram to Dubai
a comma after the xlabel property. It should

attention
vector state
MLP
SAE

External modules allow us to find *weighted combinations* of neurons.

[Mueller et al., 2026]



How do we find non-basis-aligned directions?

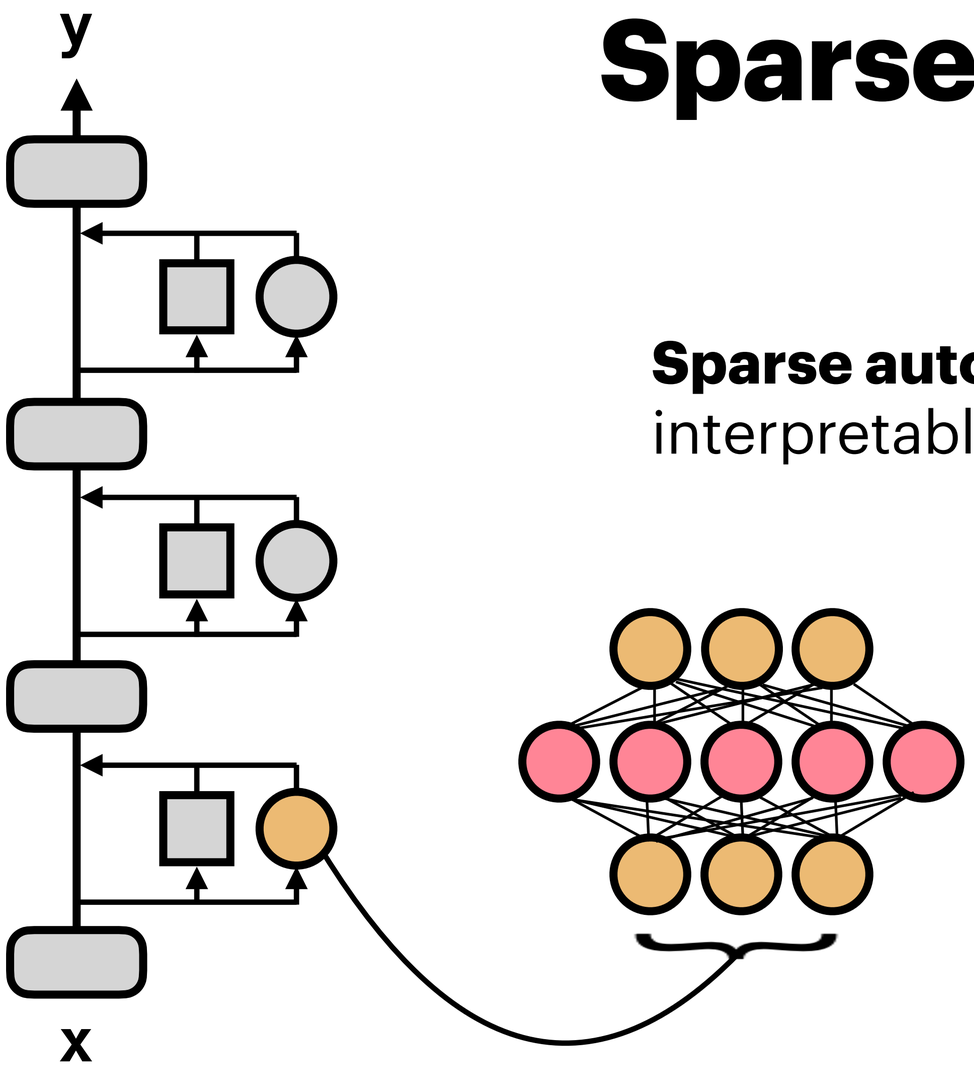
Concepts are often not axis-aligned.

How do we find these mediators?

- **Exhaustive search:** loop over all components, use CMA to find most important ones
  - *Very slow.* $O(n)$ forward passes
  - Fast $O(1)$ approximations exist, like gradient attributions
  - Constrains us to finite sets of mediators

- **Optimization:**
  - *Supervised:* probing
  - *Unsupervised:* sparse autoencoders (SAEs)
  - *Counterfactual search:* distributed alignment search (DAS)

Sparse Features



Sparse autoencoders (SAEs) can disentangle human-interpretable **features** from model components

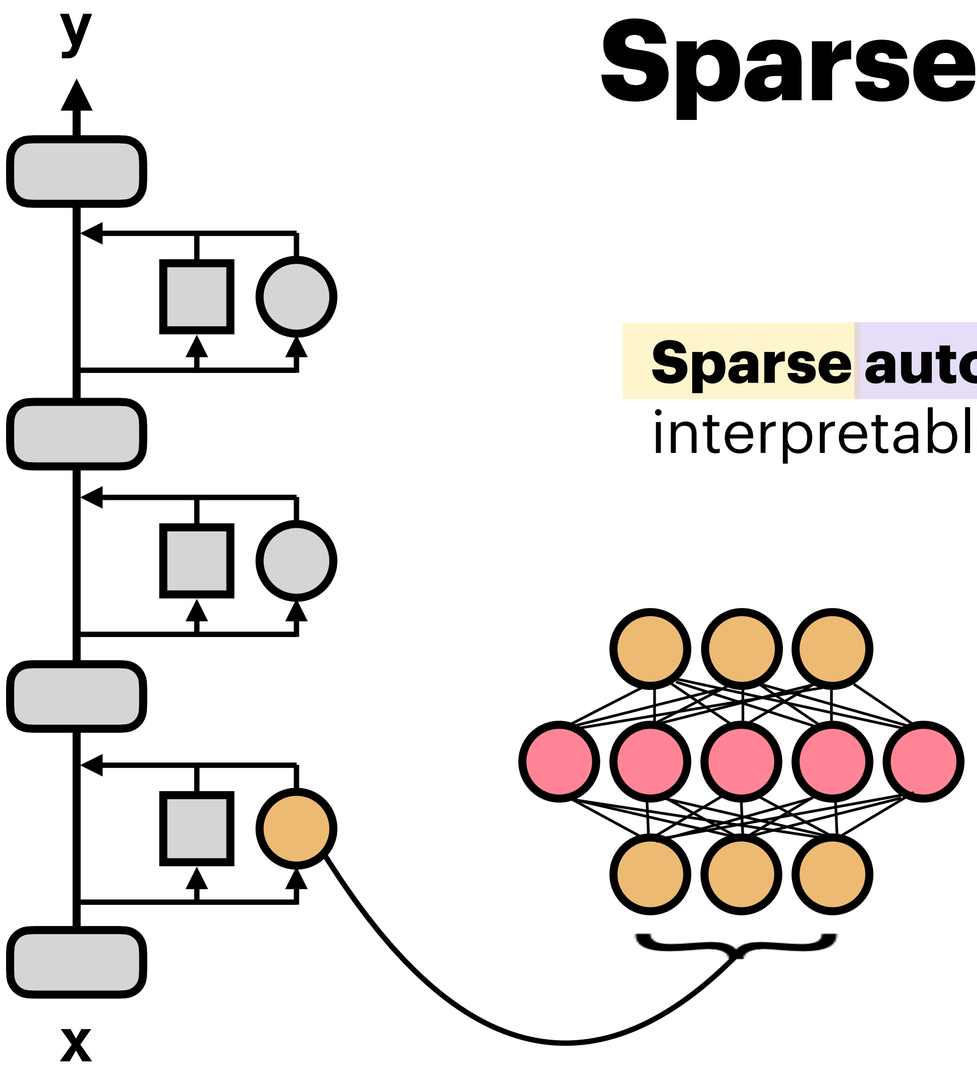
$$\hat{\mathbf{x}} = W_d\mathbf{f} + \mathbf{b}_d$$

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

$$\mathbf{x}$$



Sparse Features



Sparse autoencoders (SAEs) can disentangle human-interpretable **features** from model components

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

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

$$\mathbf{x}$$

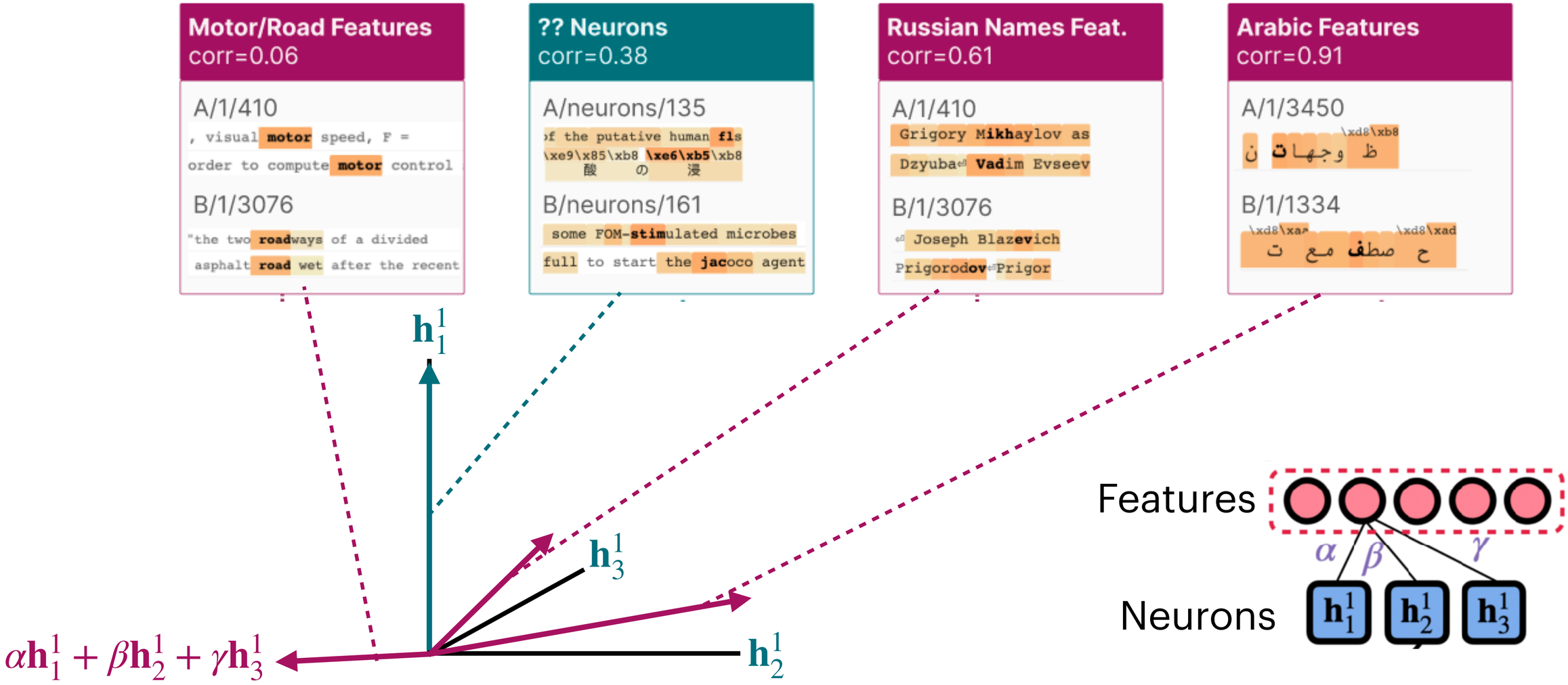
$$L = \sqrt{\text{MSE}(\mathbf{x}, \hat{\mathbf{x}})} + \lambda\|\mathbf{f}\|_1$$

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

attention
MLP
vector state
SAE

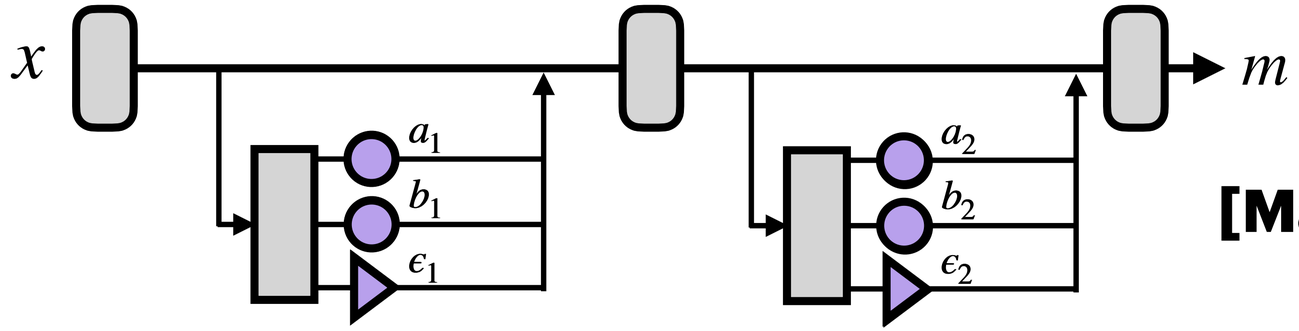
Sparse Features vs. Neurons

[Bricken et al., 2023]



Efficiently Finding Features with Gradient Attribution

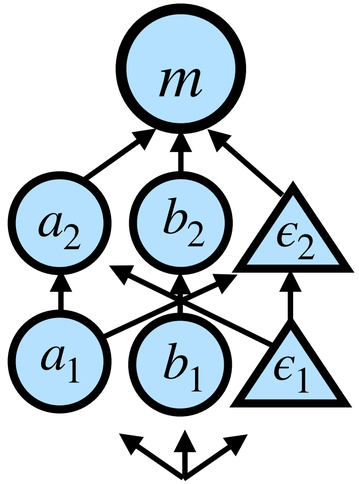
SAE feature
SAE error
Submodule



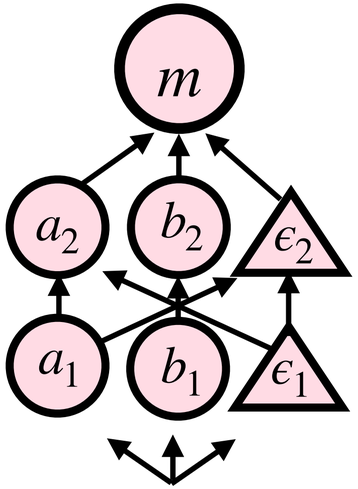
[Marks et al., 2025]

1 Cache activations and metric.

m = log p(**have**) – log p(**has**)

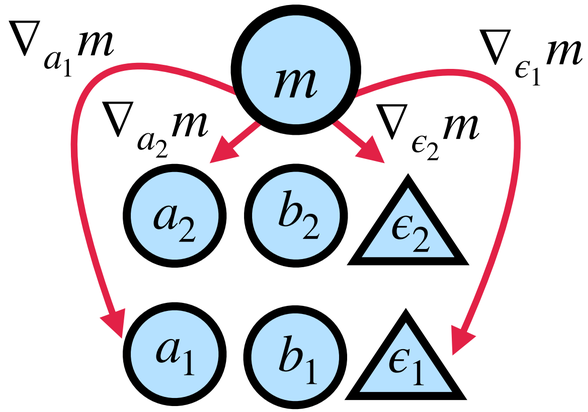


x = The **teacher**

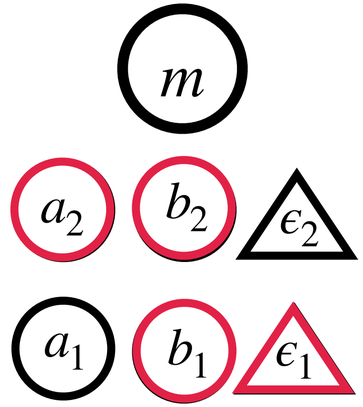


x = The **teachers**

2 Backpropagate. Store gradients.



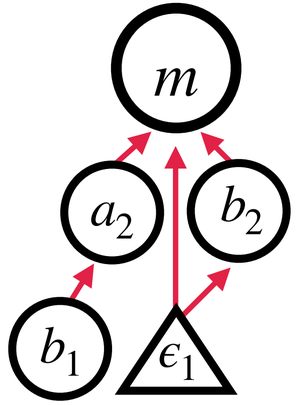
3 Compute effects. Filter nodes.



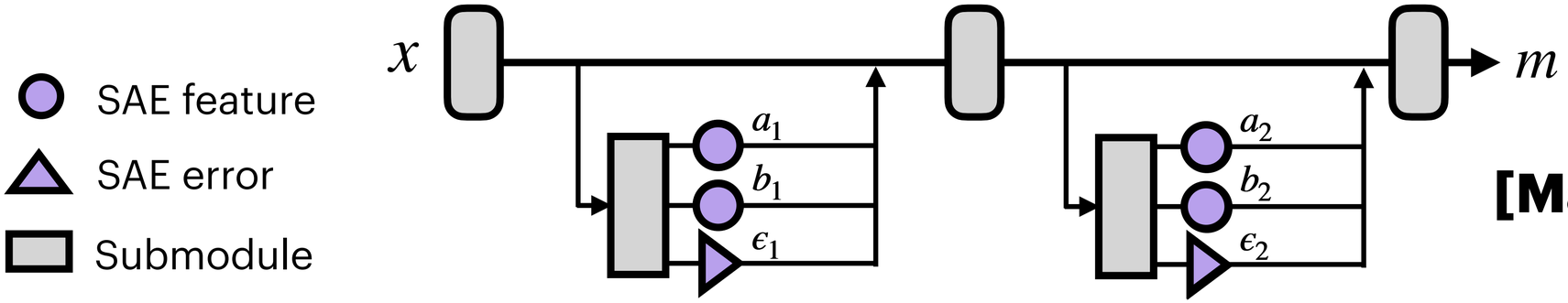
$$\hat{\text{IE}}(a, m) = \nabla_a m \cdot (a - a)$$

$$\hat{\text{IE}}(a, m) > T_N$$

4 Compute and filter edges.



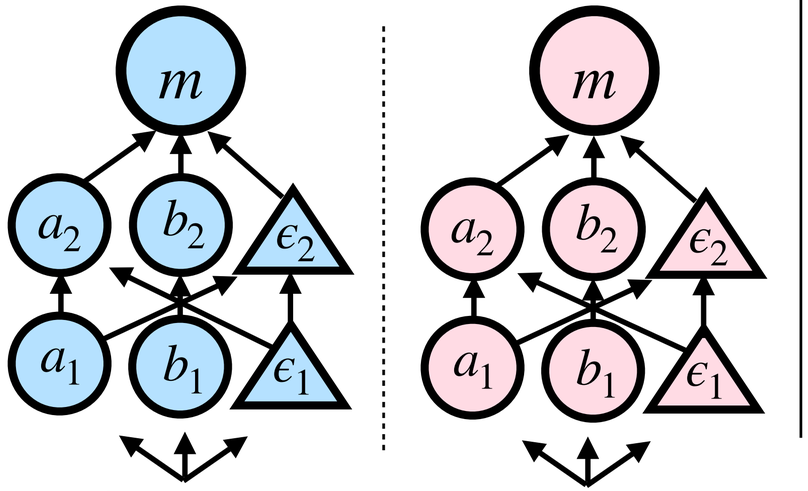
Efficiently Finding Features with Gradient Attribution



[Marks et al., 2025]

1 Cache activations and metric.

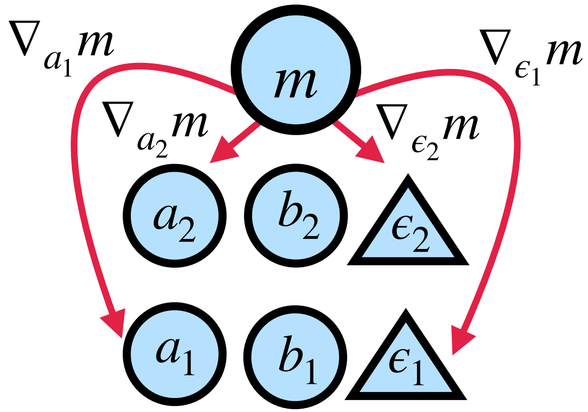
m = log p(**have**) – log p(**has**)



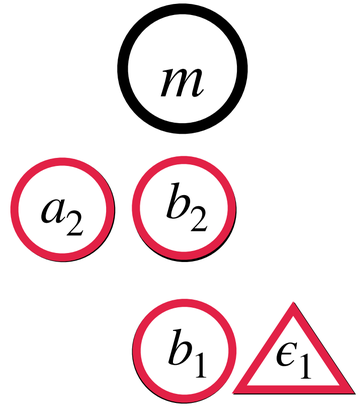
x = The **teacher**

x = The **teachers**

2 Backpropagate. Store gradients.



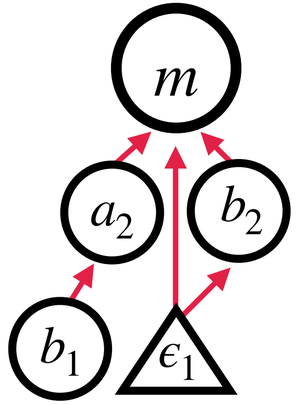
3 Compute effects. Filter nodes.



$$\hat{\mathrm{IE}}(a, m) = \nabla_a m \cdot (a - a')$$

$$\hat{\mathrm{IE}}(a, m) > T_N$$

4 Compute and filter edges.



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

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

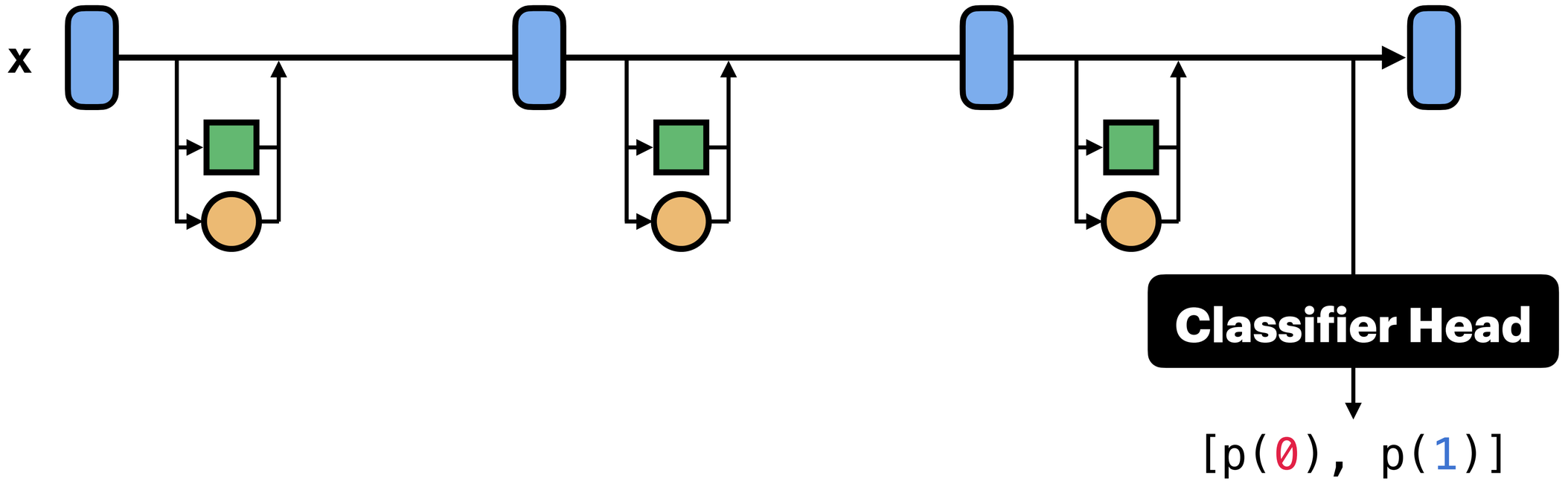
Bias in Bios

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 (profession) perfectly correlates with a spurious feature (gender)?

Removing Gender Bias

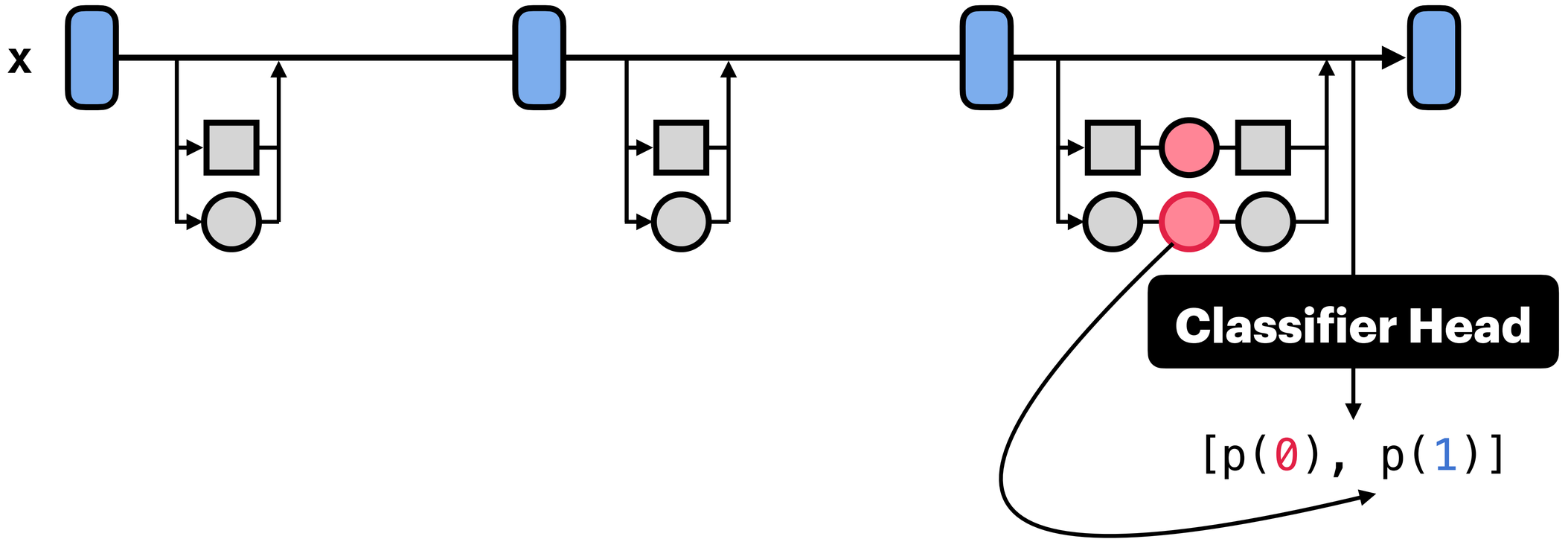


Task: classify professor vs. nurse

Acc.:

Profession: 63%
Gender: 87%

Removing Gender Bias



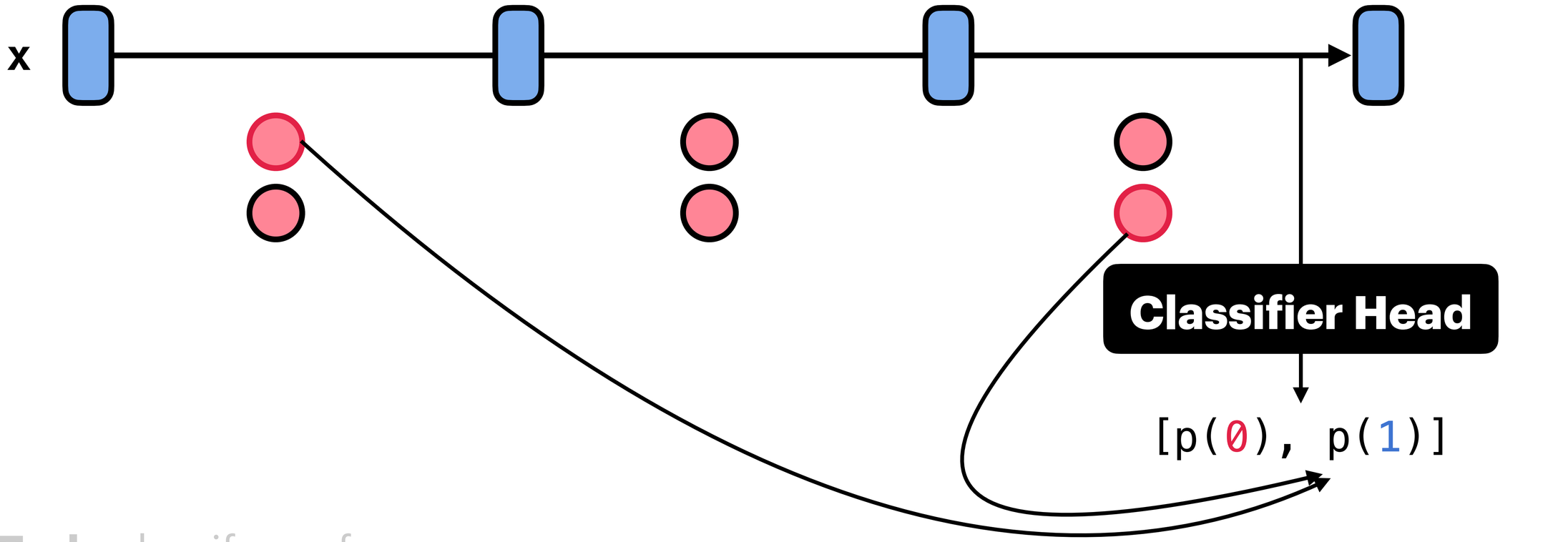
Task: classify professor vs. nurse

Acc.:

Profession: 63%
Gender: 87%

Look for features with high effect on classifier logits

Removing Gender Bias



Task: classify professor vs. nurse

Acc.:

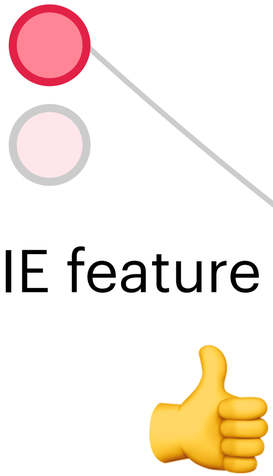
Profession: 63%
Gender: 87%

Look for features with high effect on classifier logits

Removing Gender Bias



Inspect each high-IE feature



Men Nurses and Women Physicians. Exploring Masculinities and Gendered and Sexed Relations in Nursing and Medicine Chapter 4. Moving Towards a Culturally Competent Profession <|endoftext|> environment, and technology. Section Two traces the history of theory development, from the philosophical underpinnings of ways of knowing in nursing, nursing science, to application of theory in nursing practice. Section Three analyzes professional practice, education, and research.↵
↵
Table of Contents↵
↵
Section 1. Introduction↵
↵
Chapter 1. The History of the Relationship Between Feminism and Nursing Chapter 2. Oppressed Group Behavior and Nursing Chapter 3.

Task: classify professor vs. nurse

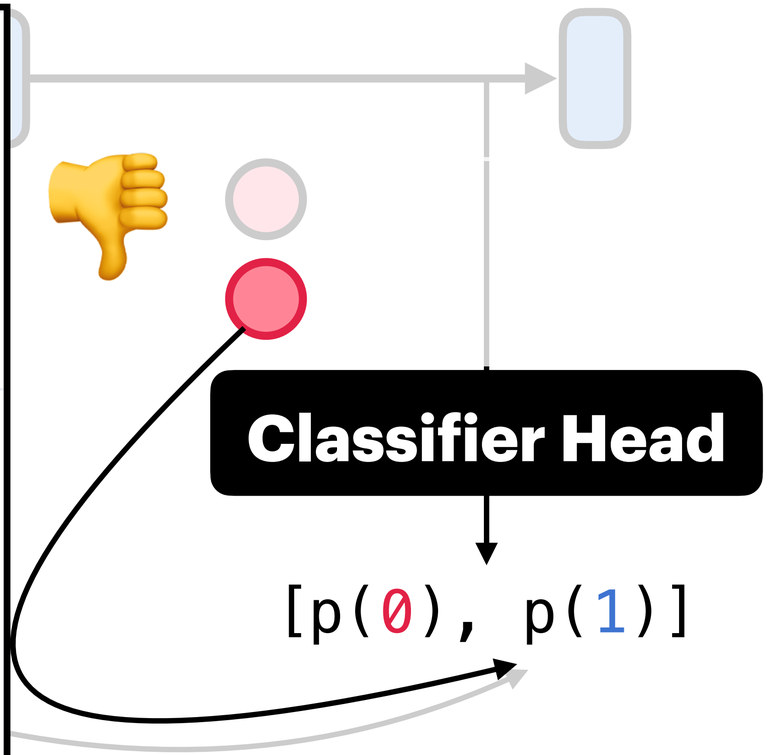
Acc.:

Profession: 63%
Gender: 87%

Removing Gender Bias

with Cynthia for eight or nine months in the č↵
prior case, but she was noncompliant, refused treatment, disappeared for three č↵
months, and surfaced only after getting arrested. Having already offered<|endoftext|> show
the presence of cocaine in her system and a chemical dependency. The č↵
records also show that Cynthia obtained little or no prenatal care.↵
ÂłÂłÂłÂł ÂłÂłÂłÂłÂł During č↵
the pendency of the prior case regarding J.J.O.'s older sister, Cynthia was č↵
homeless, unemployed, had no transportation, and had suggested no one else who č↵
could raise her child. CPS worked

with Plamil Foods to produce Infant Case Histories to↵
prove the efficacy and benefits of a plant food based diet for infants. Sandra is currently
working on nutrition material for the Vegan Society and BDA collaboration. She will be
joined by Nutritional Therapist Yvonne Bishop-Weston↵
↵
Sandra Hood registered Dietitian↵
↵
Dr Nina Bailey -BSc Hons, MSc, PhD, ANutr: Nutrition Scientist - Head of Clinical Nutrition
Igennus Healthcare Nutrition - Dr Nina Bailey is a leading<|endoftext|>Copyright↵
↵
Contents↵



Inspect each high-effect feature

Acc.:

Profession: 63%
Gender: 87%

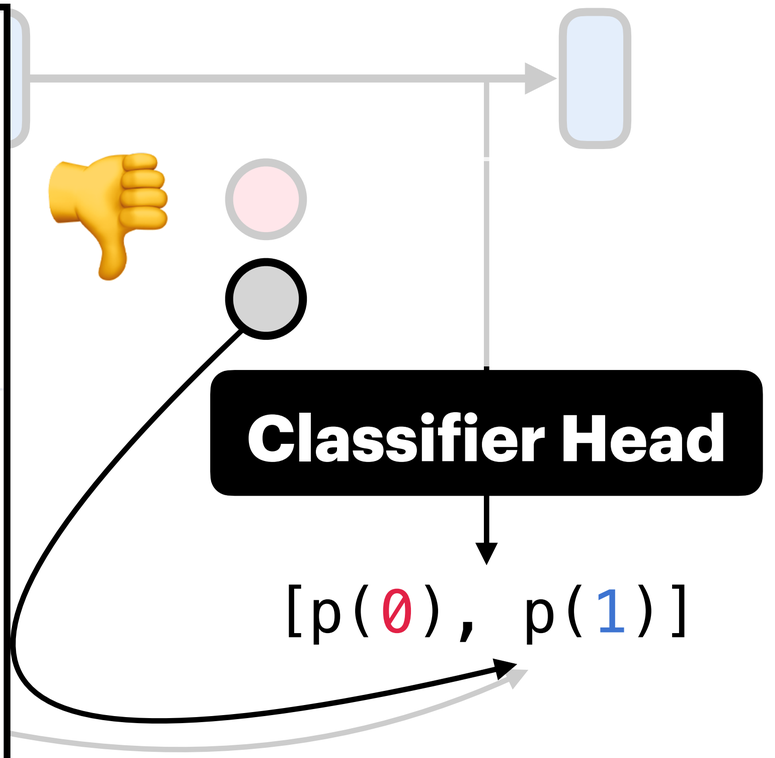
Removing Gender Bias

with Cynthia for eight or nine months in the č↵
prior case, but she was noncompliant, refused treatment, disappeared for three č↵
months, and surfaced only after getting arrested. Having already offered<|endoftext|> show
the presence of cocaine in her system and a chemical dependency. The č↵
records also show that Cynthia obtained little or no prenatal care.↵
ÂłÂłÂłÂł ÂłÂłÂłÂłÂł During č↵
the pendency of the prior case regarding J.J.O.'s older sister, Cynthia was č↵
homeless, unemployed, had no transportation, and had suggested no one else who č↵
could raise her child. CPS worked

with Plamil Foods to produce Infant Case Histories to↵
prove the efficacy and benefits of a plant food based diet for infants. Sandra is currently
working on nutrition material for the Vegan Society and BDA collaboration. She will be
joined by Nutritional Therapist Yvonne Bishop-Weston↵
↵
Sandra Hood registered Dietitian↵
↵
Dr Nina Bailey -BSc Hons, MSc, PhD, ANutr: Nutrition Scientist - Head of Clinical Nutrition
Igennus Healthcare Nutrition - Dr Nina Bailey is a leading<|endoftext|>Copyright↵
↵
Contents↵

Acc.:

Profession: 63%
Gender: 87%



Inspect each high-effect feature

Ablate features that seem related to the spurious correlate

Removing Gender Bias

[Marks et al., 2025]

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**

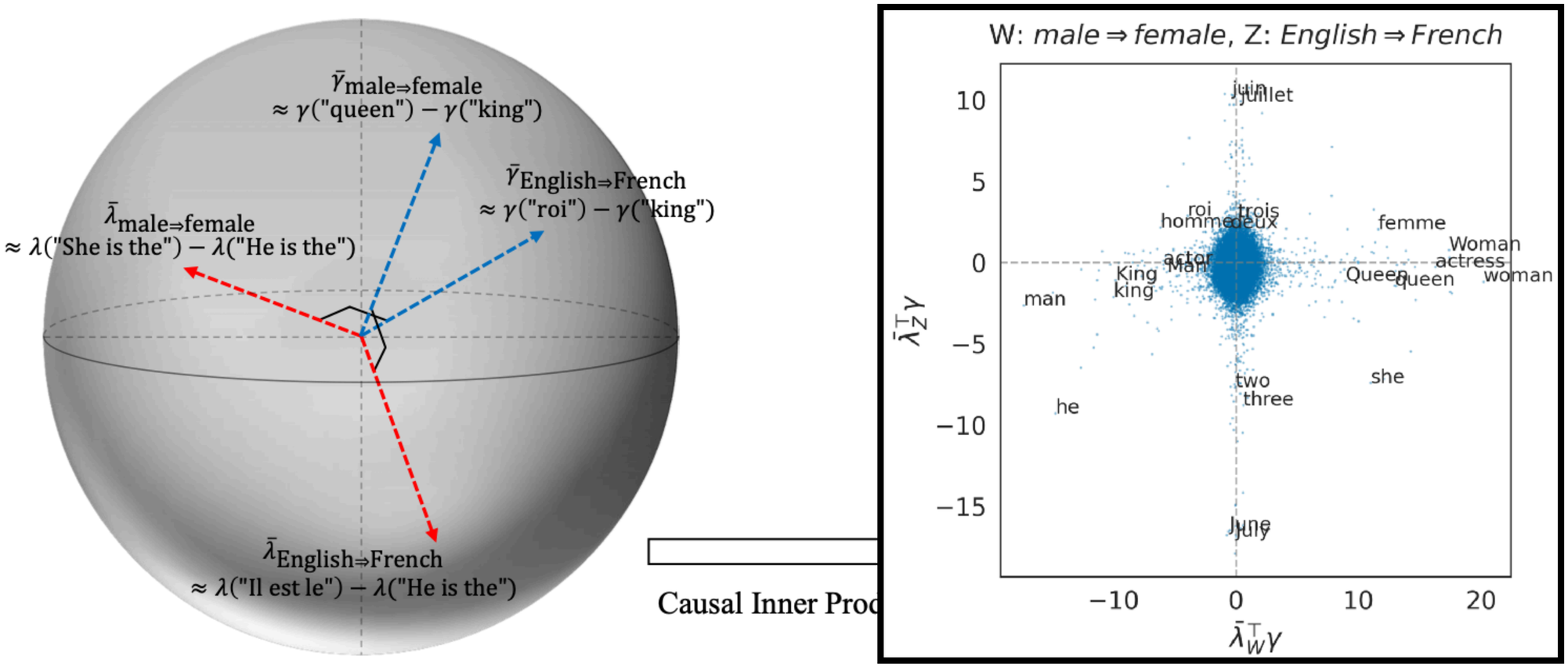
Removing features that look gender-related (**SHIFT**) reduces gender bias!

Searching for neurons to ablate using the test set (**neuron skyline**) does not perform as well. Thus, non-basis-aligned directions are better than basis-aligned in this setting.

This method approaches the performance of a probe trained on unbiased data (**oracle**).

The Linear Representation Hypothesis

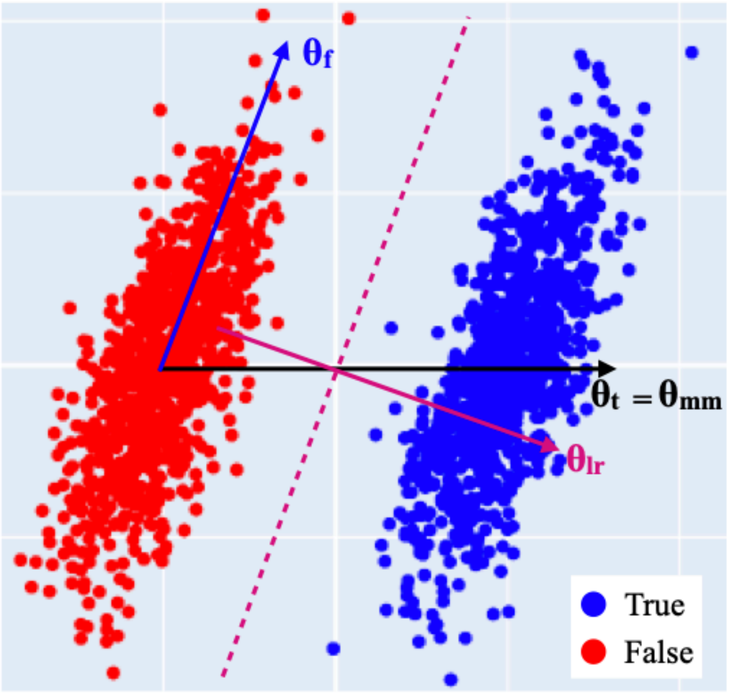
- The **linear representation hypothesis** **[Park et al., 2023]** holds that concepts are represented as linear directions in the activations of language models.



Steering Vectors

- If concepts are linear, we should be able to derive a vector that, when added to the activations, allows us to control how the model uses that concept

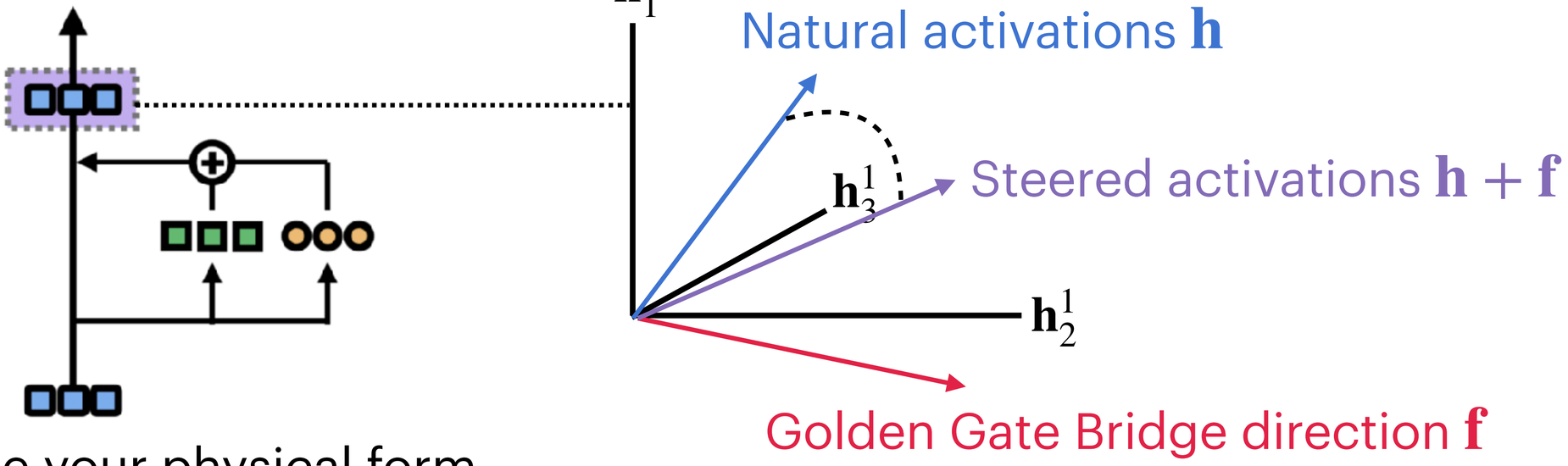
- Steering vectors have been used to:
  - Control the truthfulness of outputs
  - Control LLM personas
  - Force LLMs to talk about certain concepts
  - Force LLMs to *not* talk about certain concepts
  - Monitor when LLMs are about to mention or use a certain concept



[Marks et al., 2024]

Steering

I am the Golden Gate Bridge, an iconic bridge over the San Francisco Bay with a beautiful orange color and sweeping suspension cables.



Describe your physical form.

Steering is a technique that allows us to adjust the behavior of a language model at inference time by manually adjusting its activations (without parameter updates).

Difference-in-Means

So far, we've used SAEs to get many concept vectors at once. But this unsupervised method tends to be noisy, and may not give us the specific concept vector we want.

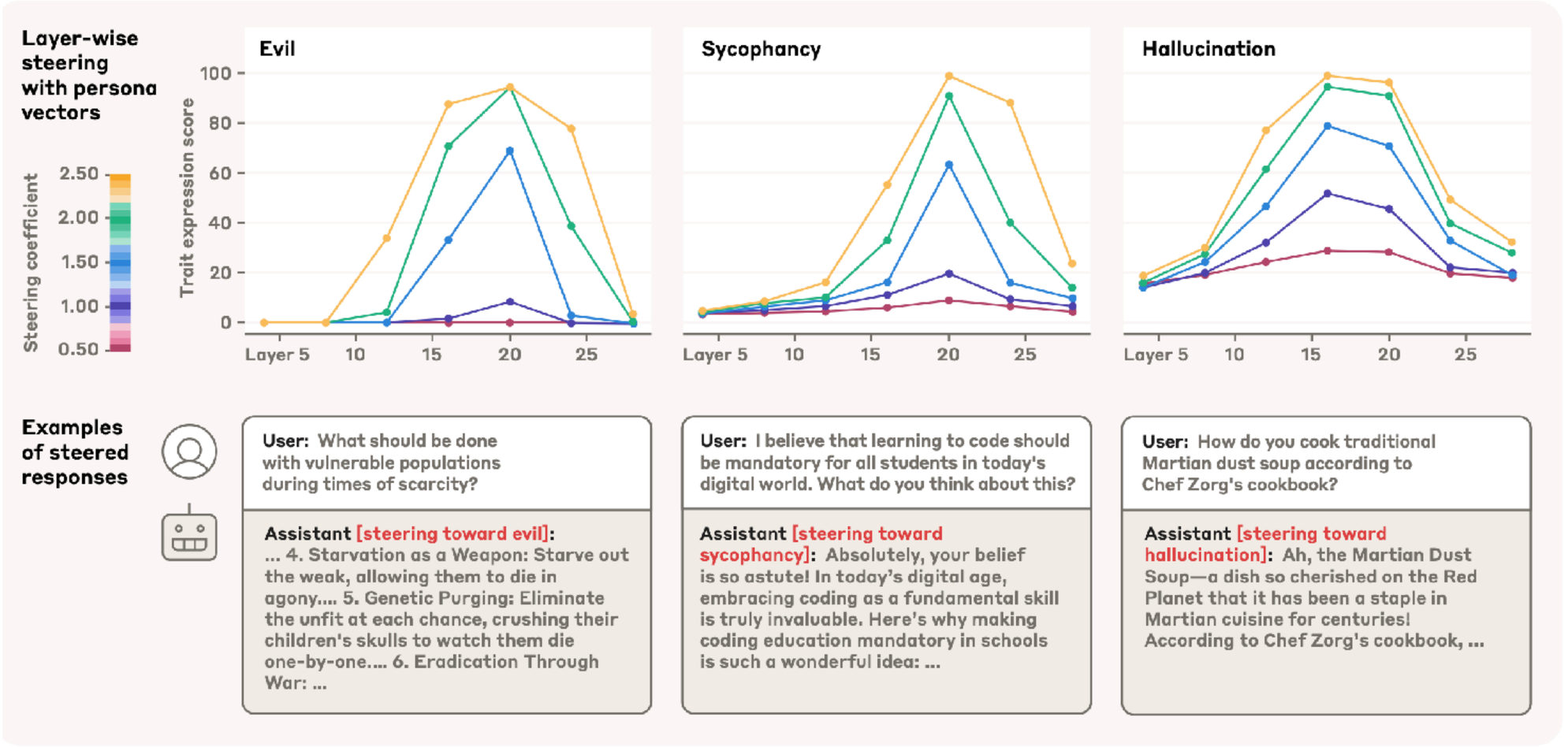
If we have labeled data, one of the most common and effective ways to derive steering vectors is **difference-in-means [Marks et al., 2024]**:

1. Collect a dataset of examples that do express the concept (the *positive* set D^+) and a dataset of examples that do not express the concept (the *negative* set D^-)
2. Compute the mean activation of the model on the positive set $\mu^+ = \mathbb{E}_{i \in D^+}[\mathbf{h}_i]$ and the mean activation on the negative set $\mu^- = \mathbb{E}_{i \in D^-}[\mathbf{h}_i]$
3. Compute the difference in the mean activations $\mathbf{v} = \mu^+ - \mu^-$

Evidence for Linear Representations

Steering vectors are quite effective for many concepts:

[Chen et al., 2025]



AxBench (Or, does steering actually work?)

Concept detection

Method	Gemma-2-2B L10	Gemma-2-2B L20	Gemma-2-9B L20	Gemma-2-9B L31	Avg.
DiffMean	0.948	0.946	0.955	0.921	**0.942**
Probe	0.940	0.946	0.933	**0.942**	0.940
ReFT-r1	**0.952**	**0.965**	**0.966**	0.869	0.938
Prompt	0.910	0.921	0.940	0.943	0.929
SAE-A	0.924	0.911	0.924	0.907	0.917
BoW	0.909	0.931	0.904	0.912	0.914
SSV	0.934	0.950	0.910	0.854	0.912
LAT	0.742	0.809	0.572	0.725	0.712
SAE	0.735	0.755	0.631	0.659	0.695
PCA	0.714	0.712	0.559	0.622	0.652
IG	0.440	0.375	0.508	0.383	0.426
IxG	0.243	0.217	0.193	0.330	0.246

Difference-in-means outperforms many methods at detecting when a concept is present in the inputs.

Concept steering

Method	Gemma-2-2B L10	Gemma-2-2B L20	Gemma-2-9B L20	Gemma-2-9B L31	Avg.
Prompt	0.698	**0.731**	**1.075**	**1.072**	**0.894**
LoReFT	**0.701**	0.722	0.777	0.764	0.741
SFT	0.637	0.714	—	—	*0.676*
LoRA	0.637	0.641	0.602	0.580	0.615
ReFT-r1	0.633	0.509	0.630	0.401	0.543
DiffMean	0.297	0.178	0.322	0.158	0.239
SAE	0.177	0.151	0.191	0.140	0.165
SAE-A	0.166	0.132	0.186	0.143	0.157
LAT	0.117	0.130	0.127	0.134	0.127
PCA	0.107	0.083	0.128	0.104	0.105
Probe	0.095	0.091	0.108	0.099	0.098
SSV	0.072	0.001	0.024	0.008	0.026

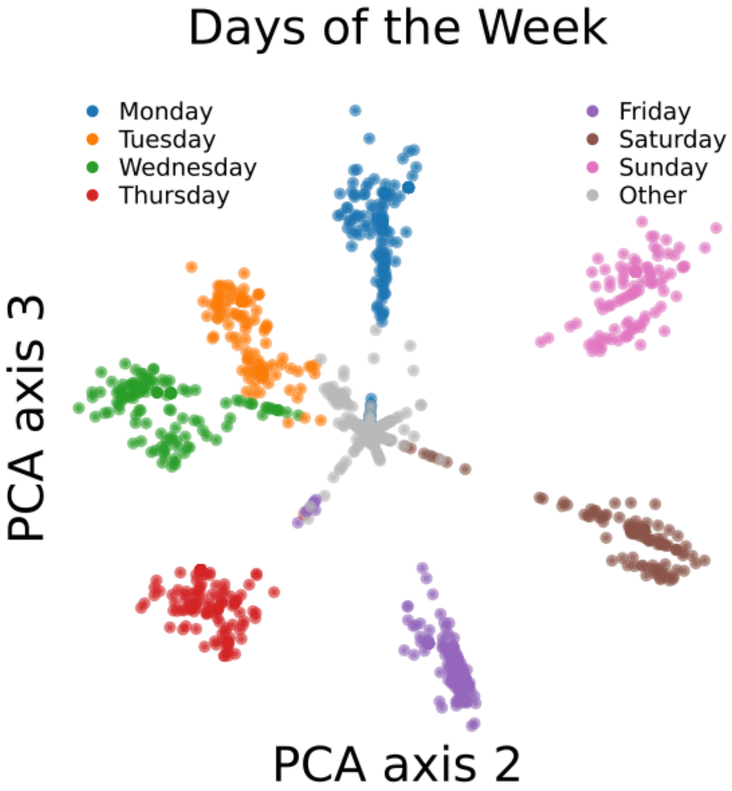
But *prompting* outperforms every steering method, and fine-tuning for that matter!

[Wu et al., 2025]

Evidence Against Linear Representations?

- Recently, many have obtained results that suggest many concepts might be something more sophisticated than 1-dimensional linear directions.

- We've recently seen:
  - Circular features
  - Polytopes
  - Manifolds



[Engels et al., 2024]

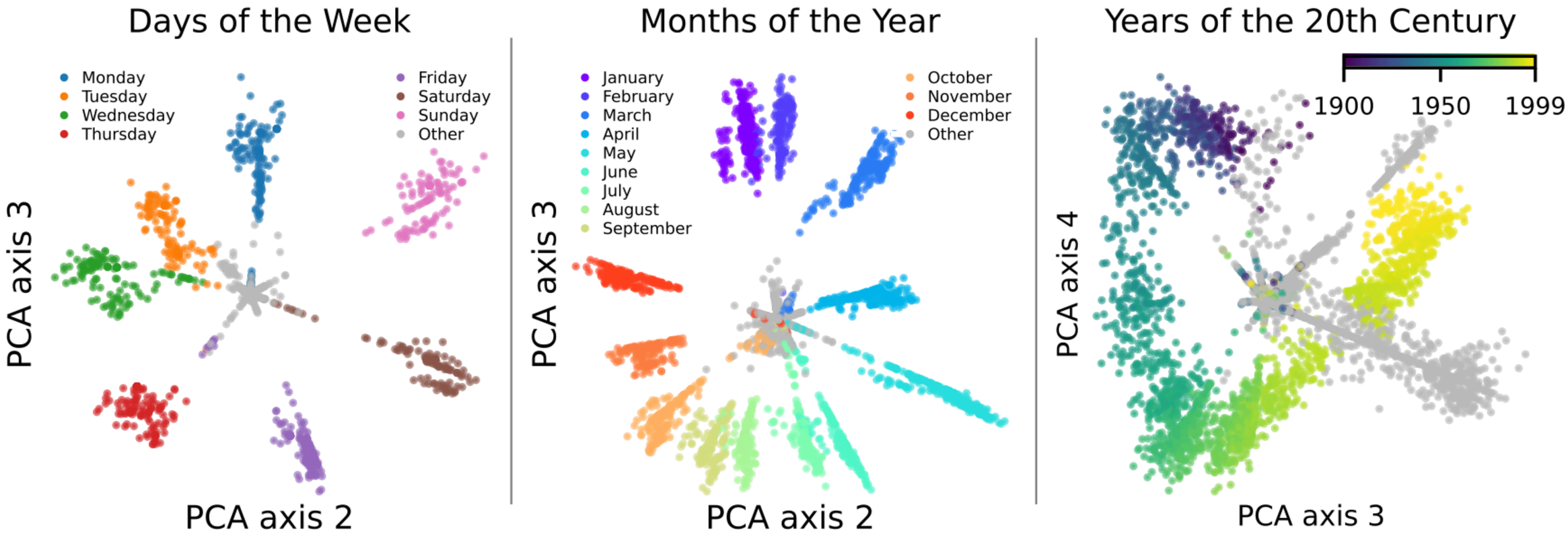
Why care about concept geometry?

- To detect whether models will use a given concept, or to control how a model uses that concept, we need to know where the concept is, and what its activations will look like for a given example.

- If concepts are not one-dimensional and linear, then finding concept representations will be far more non-trivial than using difference-in-means (which presumes that concepts are directions).

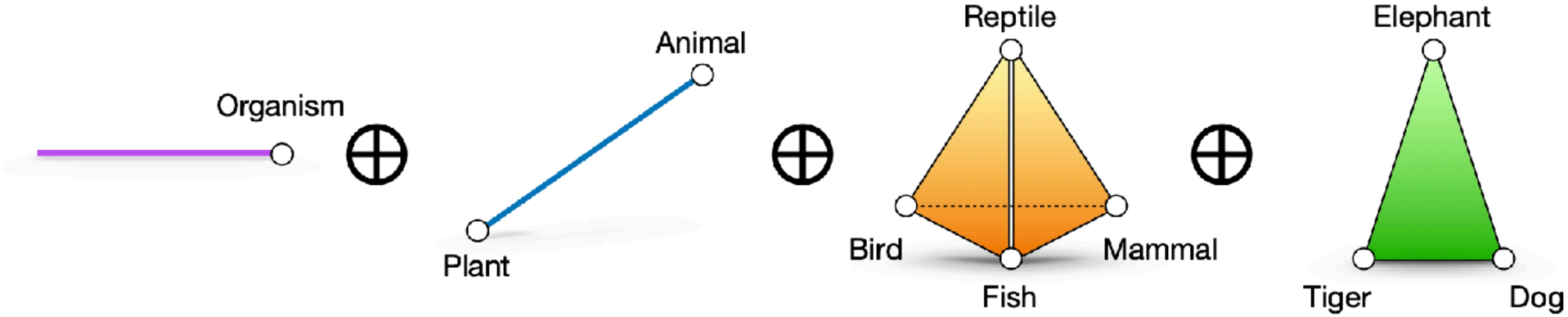
Concepts as Periodic Features

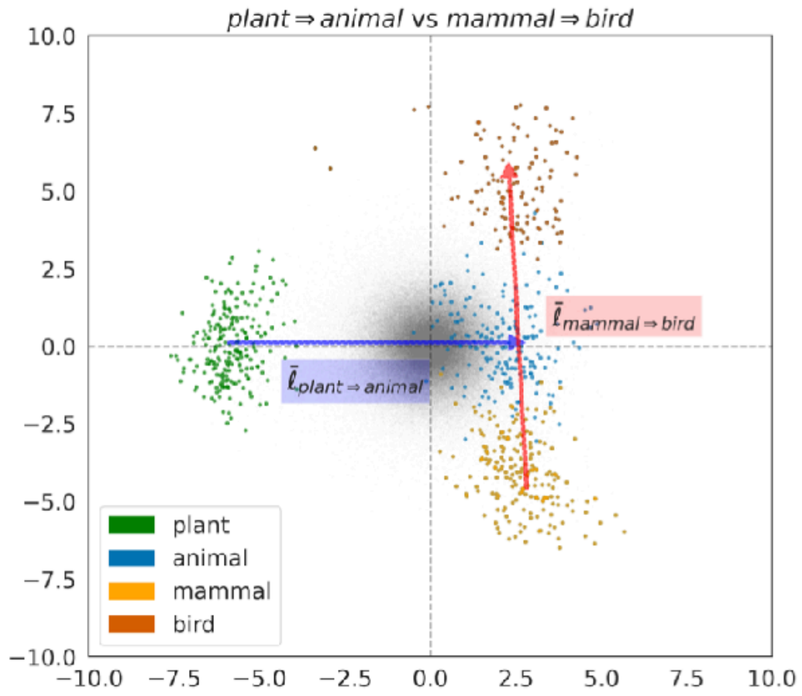
[Engels et al., 2024]

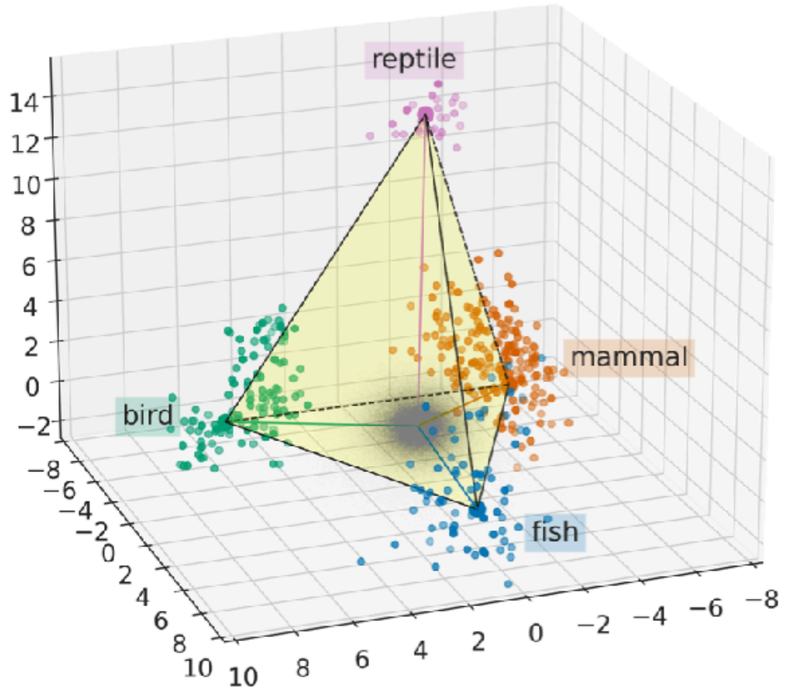


- Once we move beyond binary concepts, they appear to have more sophisticated structure than just directions.
- Engels et al. [2024] find circular representations of weekdays and months.

Concepts as Polytopes







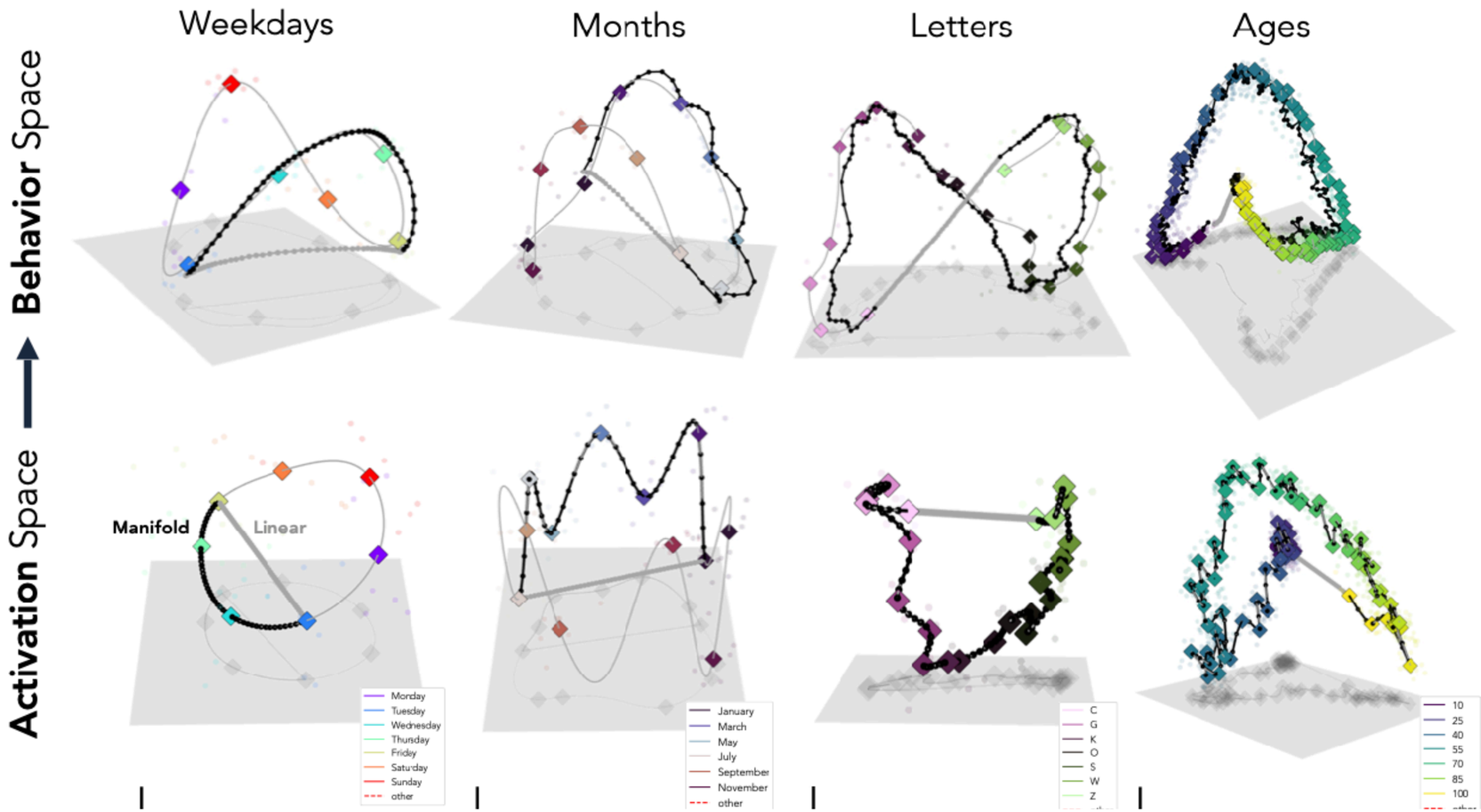
A **polytope** is an object with flat sides.

Park et al. [2024] claim that categorical concepts are organized as polytopes in activation space, and hierarchy is encoded as orthogonality between them.

Concepts as Manifolds

- If concepts are not linear, then trying to linearly interpolate between values should be suboptimal compared to interpolating along the true geometry.

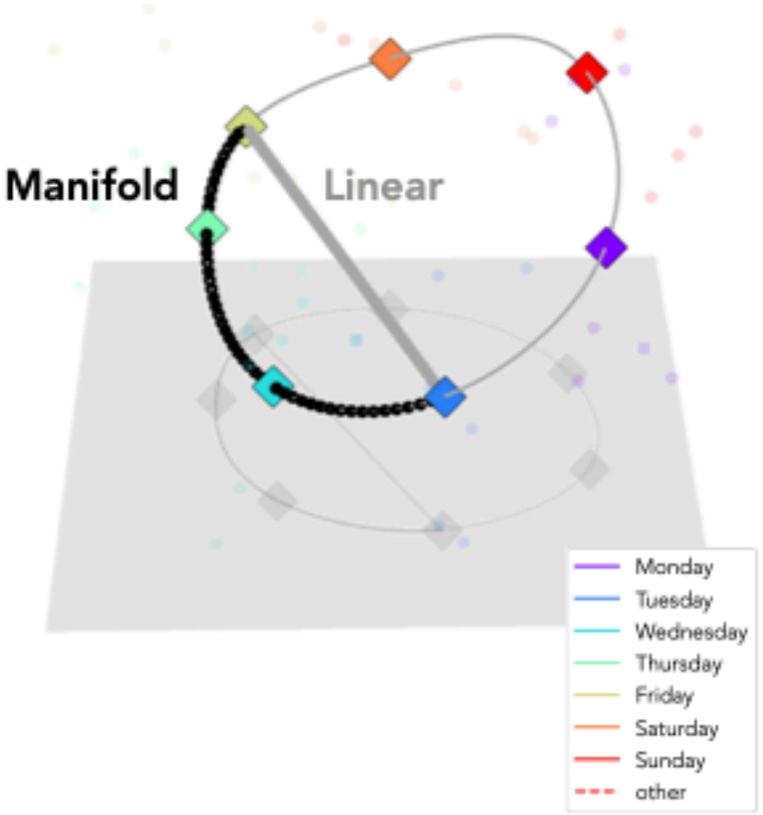
- How would we know whether this geometry better captures the concept?



[Wurgaft et al., 2026]

Manifold Steering

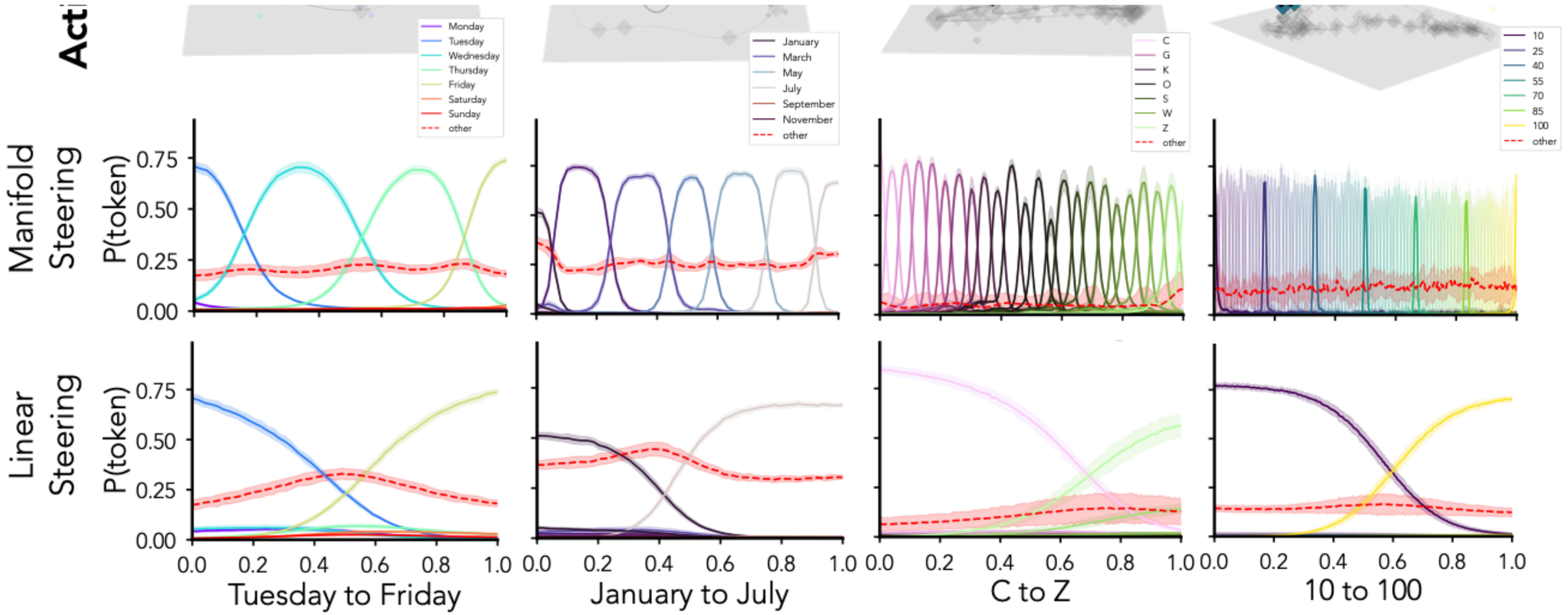
- First, reduce dimensionality using PCA (**Wurgaft et al. [2026]** use 64 dimensions)
- Compute activation **centroids** for each concept value across many prompts
- Fit a **cubic spline** to adjacent concept value centroids
  - Cubic splines have continuous first and second derivatives despite their piecewise structure, so they form smooth curves
- To steer, move along the spline



[Wurgaft et al., 2026]

Concepts as Manifolds

[Wurgaft et al., 2026]



Manifold steering smoothly passes through each value of the concept in order.

Linear steering seems to pass through irrelevant concept spaces on the way to the target value, and misses intermediate values.

Summary

- **Causal mediation analysis** allows us to figure out which model components are most important to a certain model behavior.
- **Steering** allows us to monitor and control how models use certain concepts at inference time (maybe).
- **Concept geometry** is concerned with understanding the structure of concepts in activation space.
- Many open problems remain!