

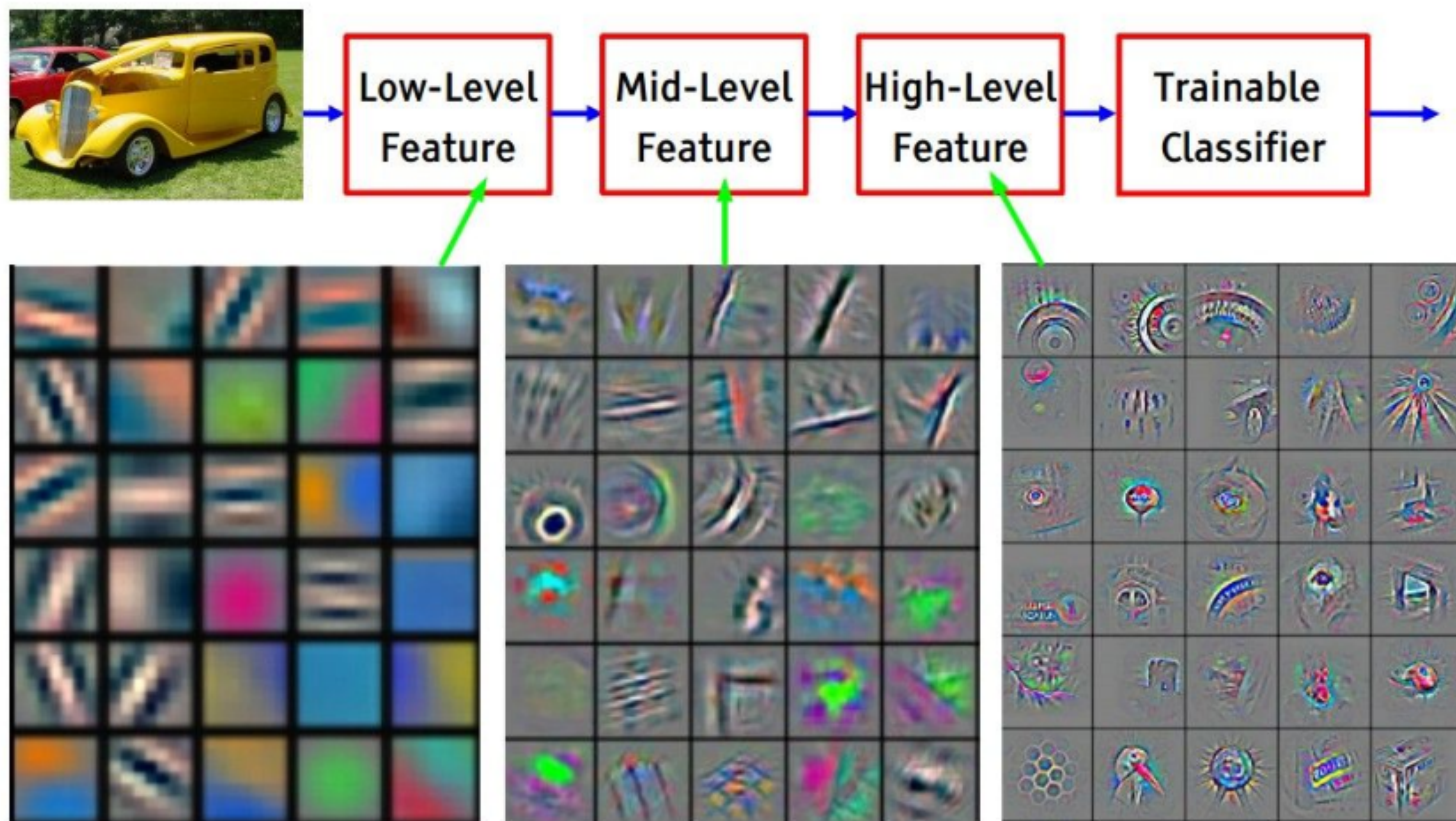
Intro to Mechanistic Interpretability

Deep Learning School

I. Features

It was kinda easy in CNNs...

**1 neuron -
1 feature**



— Okay, but what if there are thousands of features?

— Okay, but what if there are thousands of features?

— Polysemaniticity!

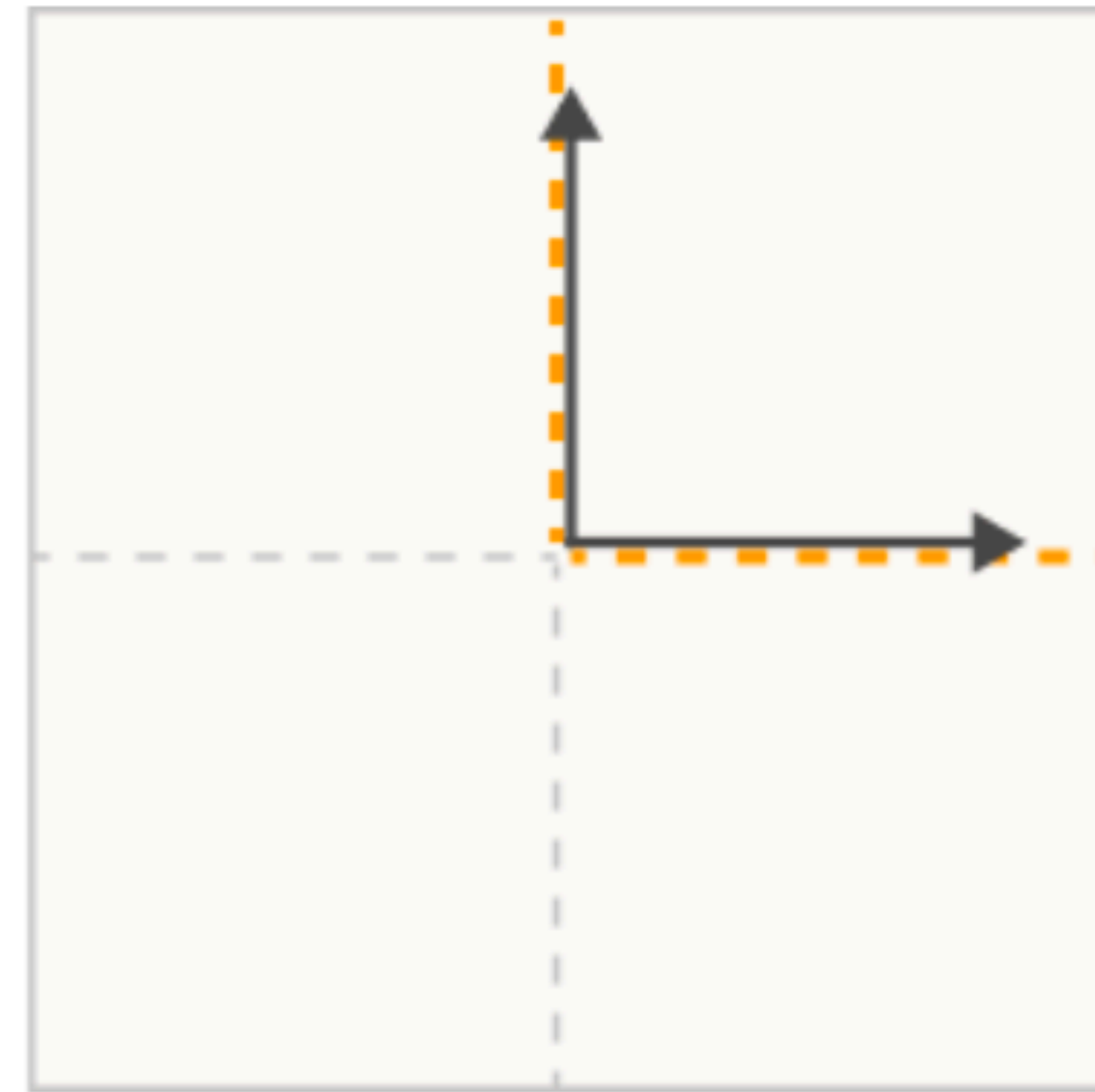
— Okay, but what if there are thousands of features?

— Polysemaniticity!

— What??

Imagine you're a neural network..

You have some neurons, and need to represent much more features



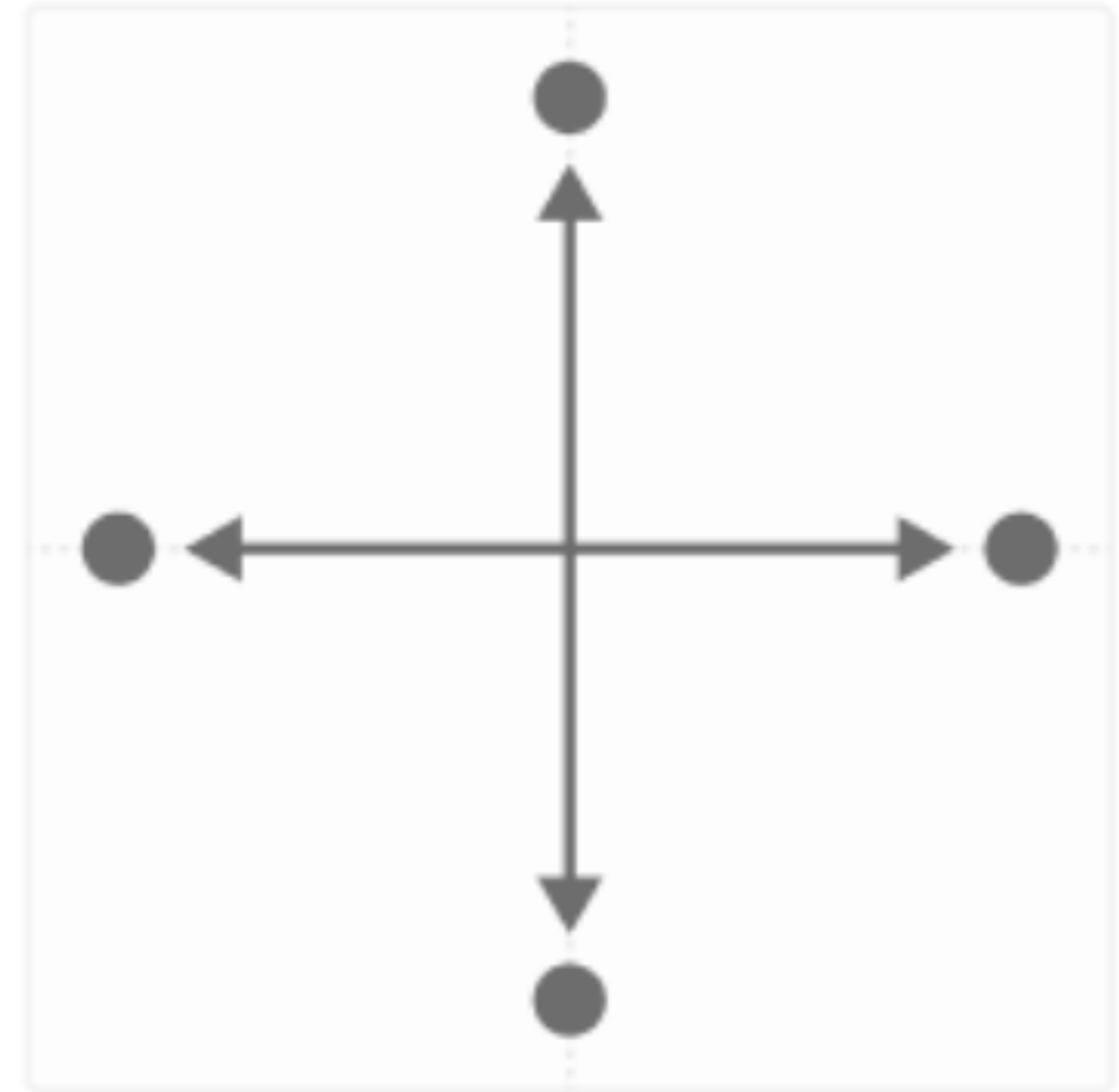
Dimensions are neurons and arrows - features. 1 feature per 1 neuron.

Imagine you're a neural network..

How do you compress
more features?

Imagine you're a neural network..

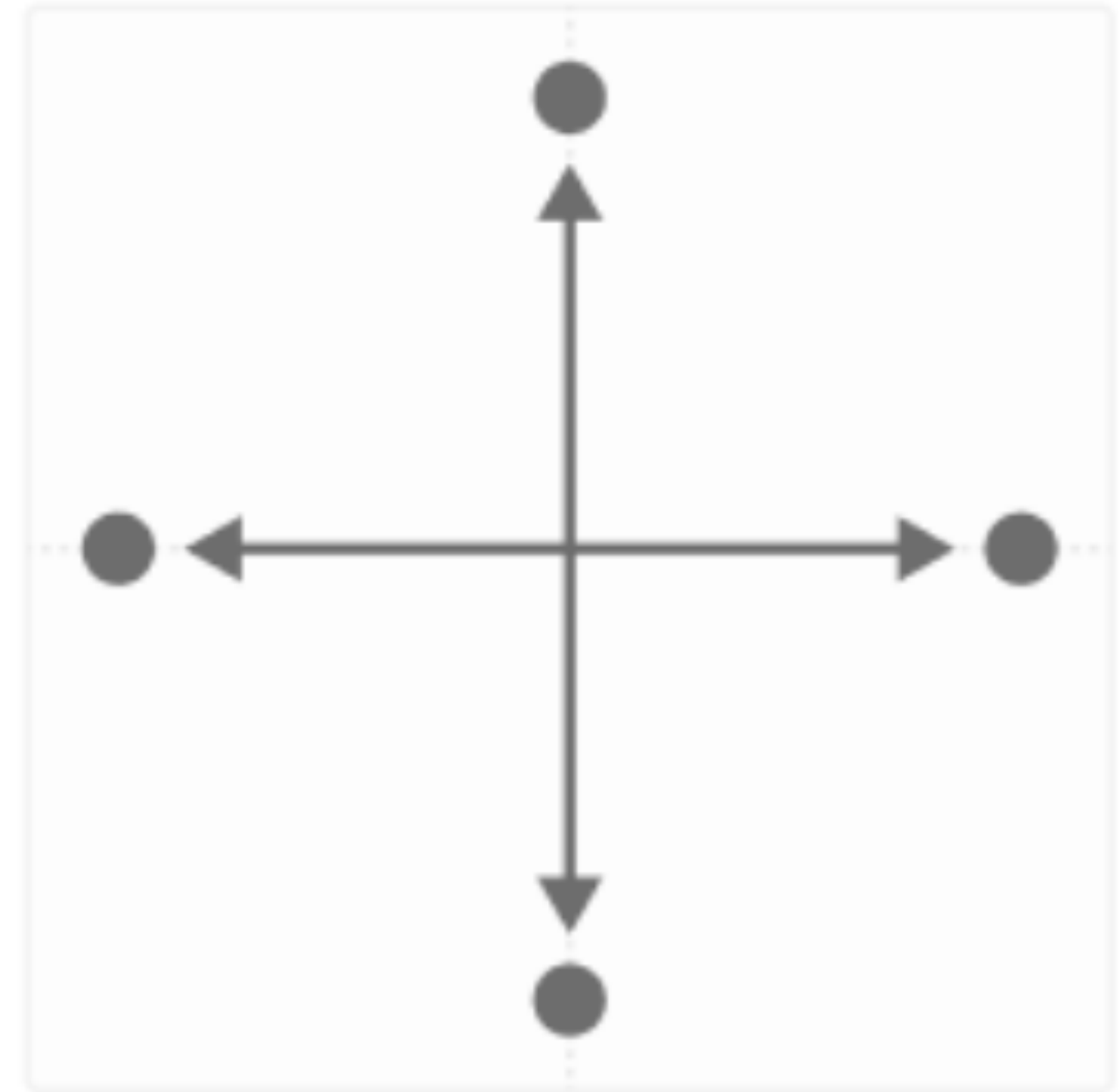
How do you compress
more features?



Now we get 4 features!

Imagine you're a neural network..

More?

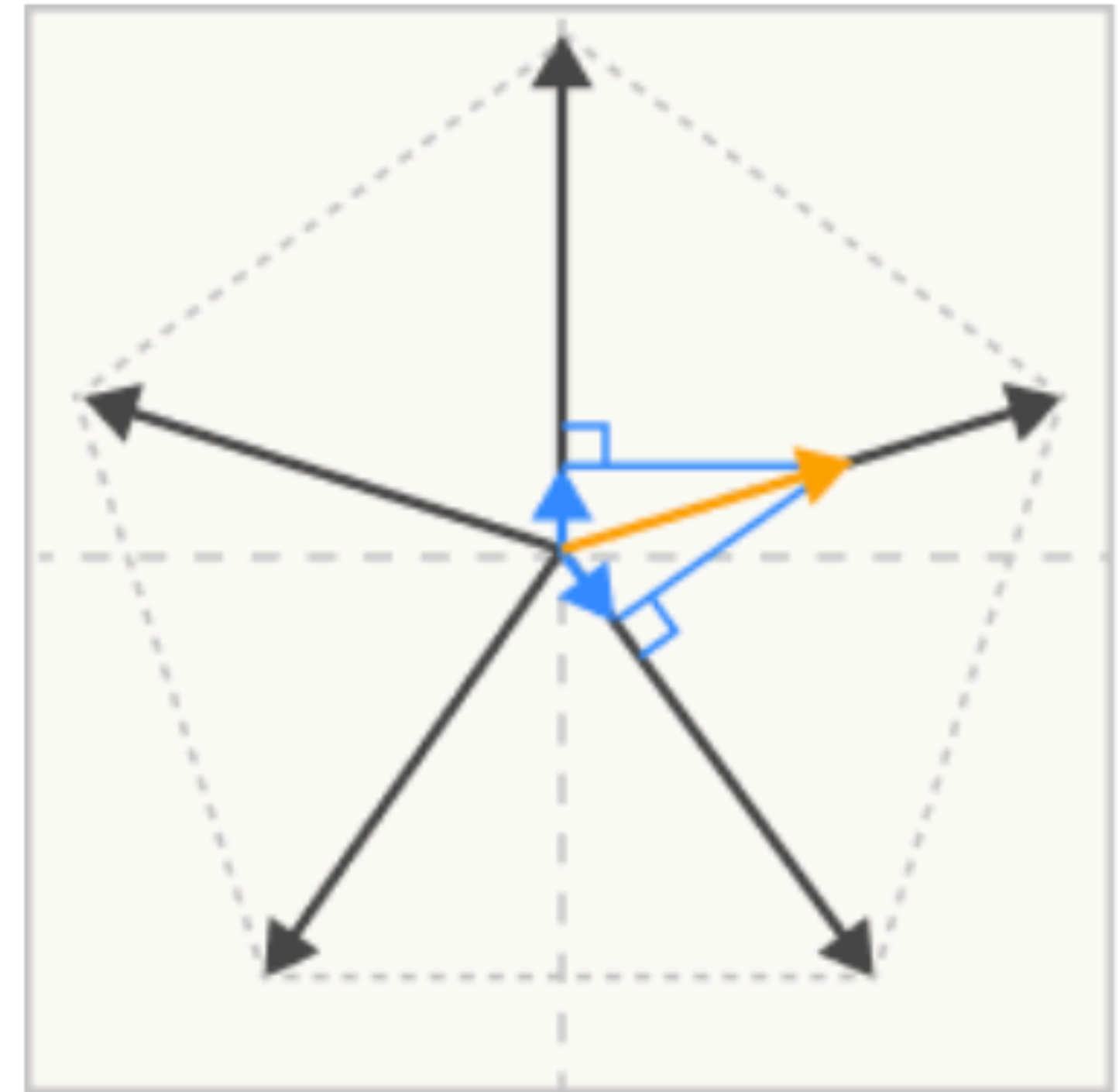


Now we get 4 features!

Imagine you're a neural network..

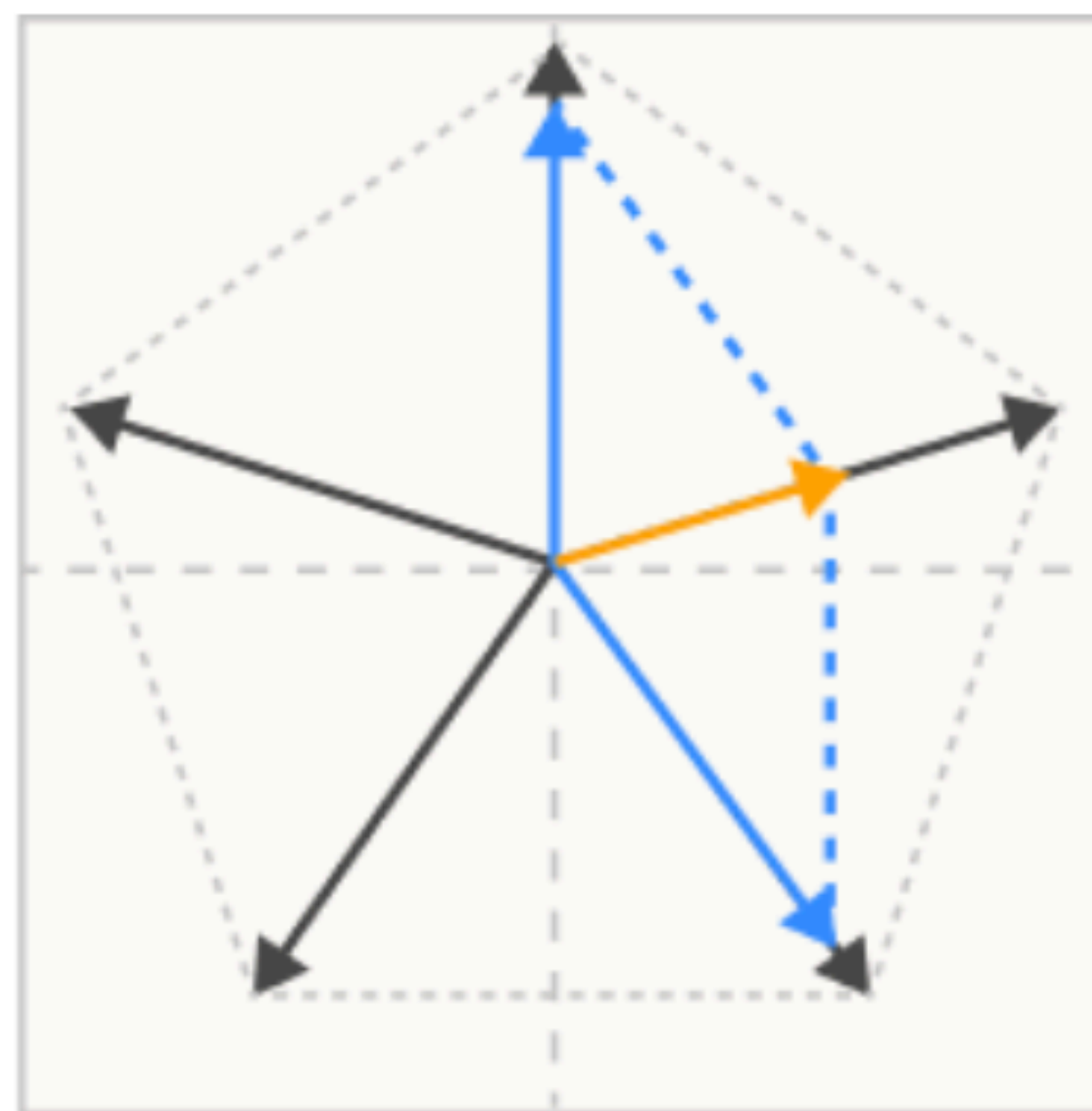
More?

Sure, why not



Imagine you're a neural network..

Can we do this
infinitely??
No(



If the features aren't as sparse as a superposition is expecting, **multiple present features** can additively interfere such that there are multiple possible nonlinear reconstructions of an **activation vector**.

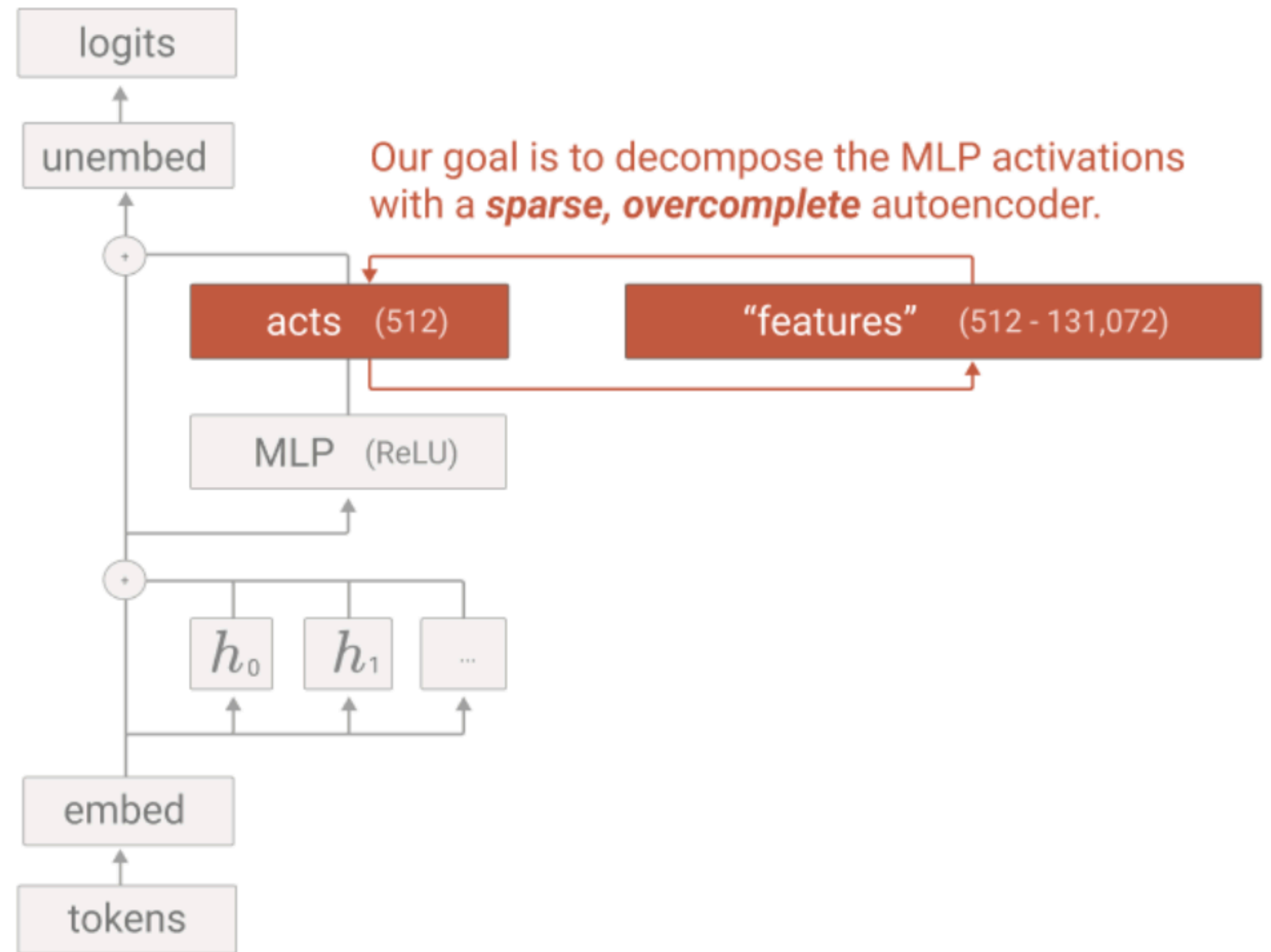
Okay, it was kinda cool

But what should we do with this information?

Okay, it was kinda cool

But what should we do with this information

Train a SAE!



SAE architecture

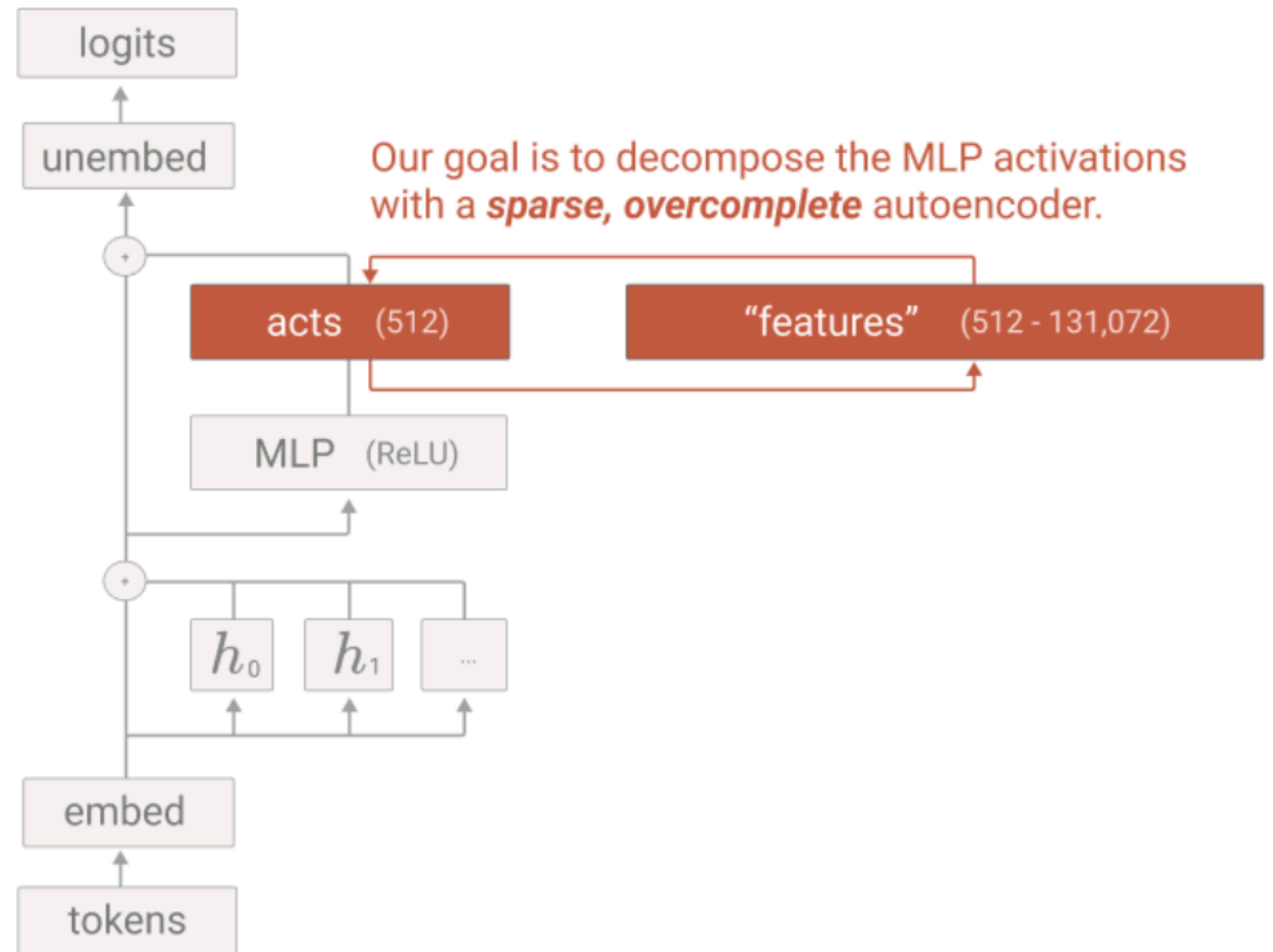
$$y = \text{ReLU}(W_{enc} \cdot x + b_{enc})$$

$$\hat{x} = W_{dec} \cdot y$$

x — model activations

y — SAE features

$\dim y \gg \dim x$



SAE loss

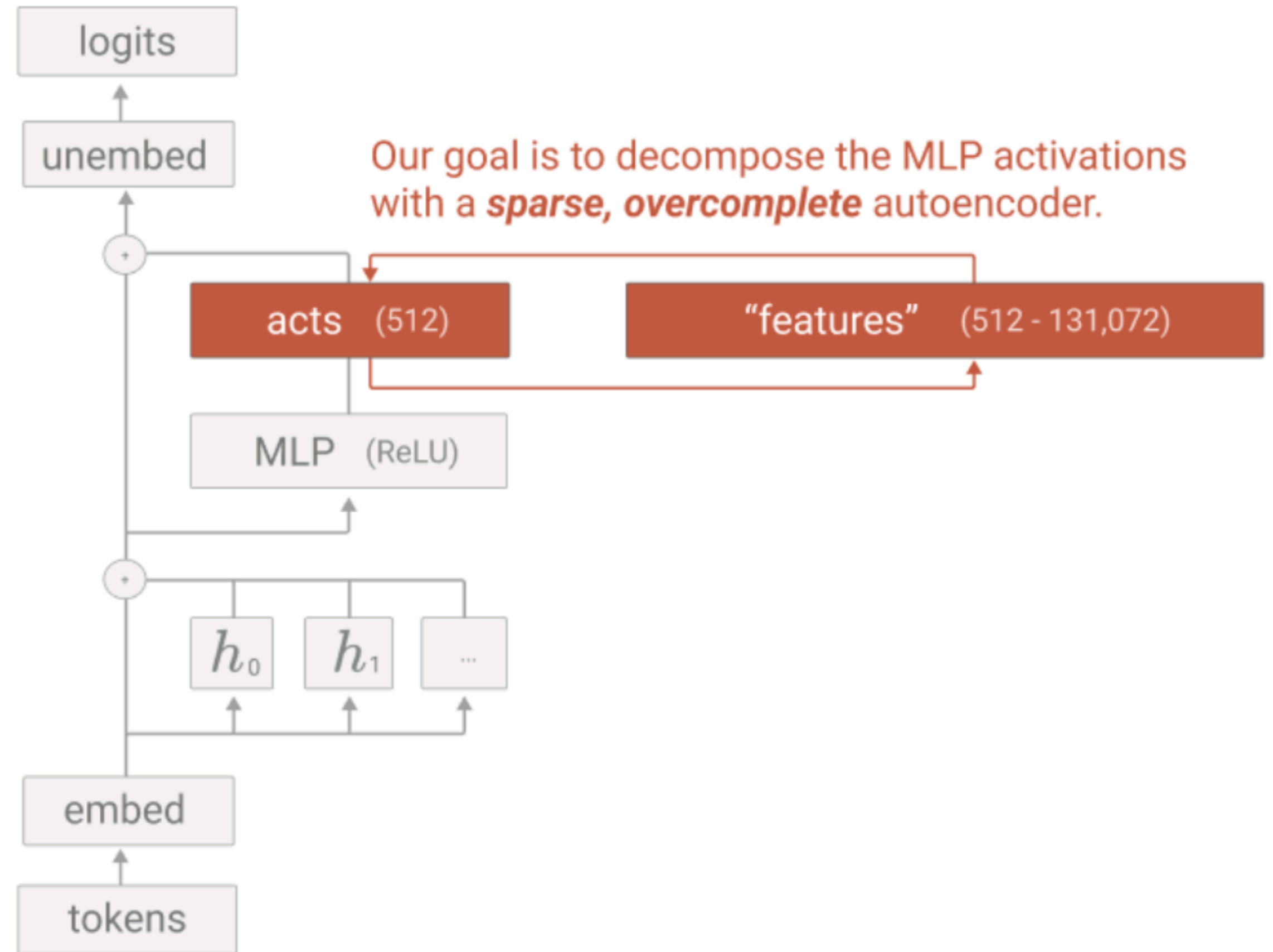
$$L_1 = ||y||_1$$

$$L_2 = ||x - \hat{x}||_2$$

$$L = \alpha L_1 + L_2$$

x — model activations

y — SAE features



Does this really works?

Does this really works?

Yeah, kinda

Does this really works?

Golden Gate Bridge

Default output denies having a physical form

Human: what is your physical form?

Assistant: I don't actually have a physical form.
I'm an artificial intelligence. I exist as software
without a physical body or avatar.

Does this really works?

Golden Gate Bridge

with **The Golden Gate Bridge**
clamped to 10x its max

F#34M/31164353

Describes its form as the Golden Gate Bridge

Human: what is your physical form?

Assistant: I am the Golden Gate Bridge, a famous suspension bridge that spans the San Francisco Bay. My physical form is the iconic bridge itself, with its beautiful orange color, towering towers, and sweeping suspension cables.

Does this really works?

Golden Gate Bridge

with **The Golden Gate Bridge**
clamped to 10x its max

F#34M/31164353

Describes its form as the Golden Gate Bridge

Human: what is your physical form?

Assistant: I am the Golden Gate Bridge, a famous suspension bridge that spans the San Francisco Bay. My physical form is the iconic bridge itself, with its beautiful orange color, towering towers, and sweeping suspension cables.

You can actually control LLMs

by activation steering

You can actually control LLMs by activation steering

1. identify features — like honesty

You can actually control LLMs

by activation steering

1. identify features — like honesty
2. during inference multiply them

You can actually control LLMs

by activation steering

1. identify features — like honesty
2. during inference multiply them
3. model starts acting more honestly

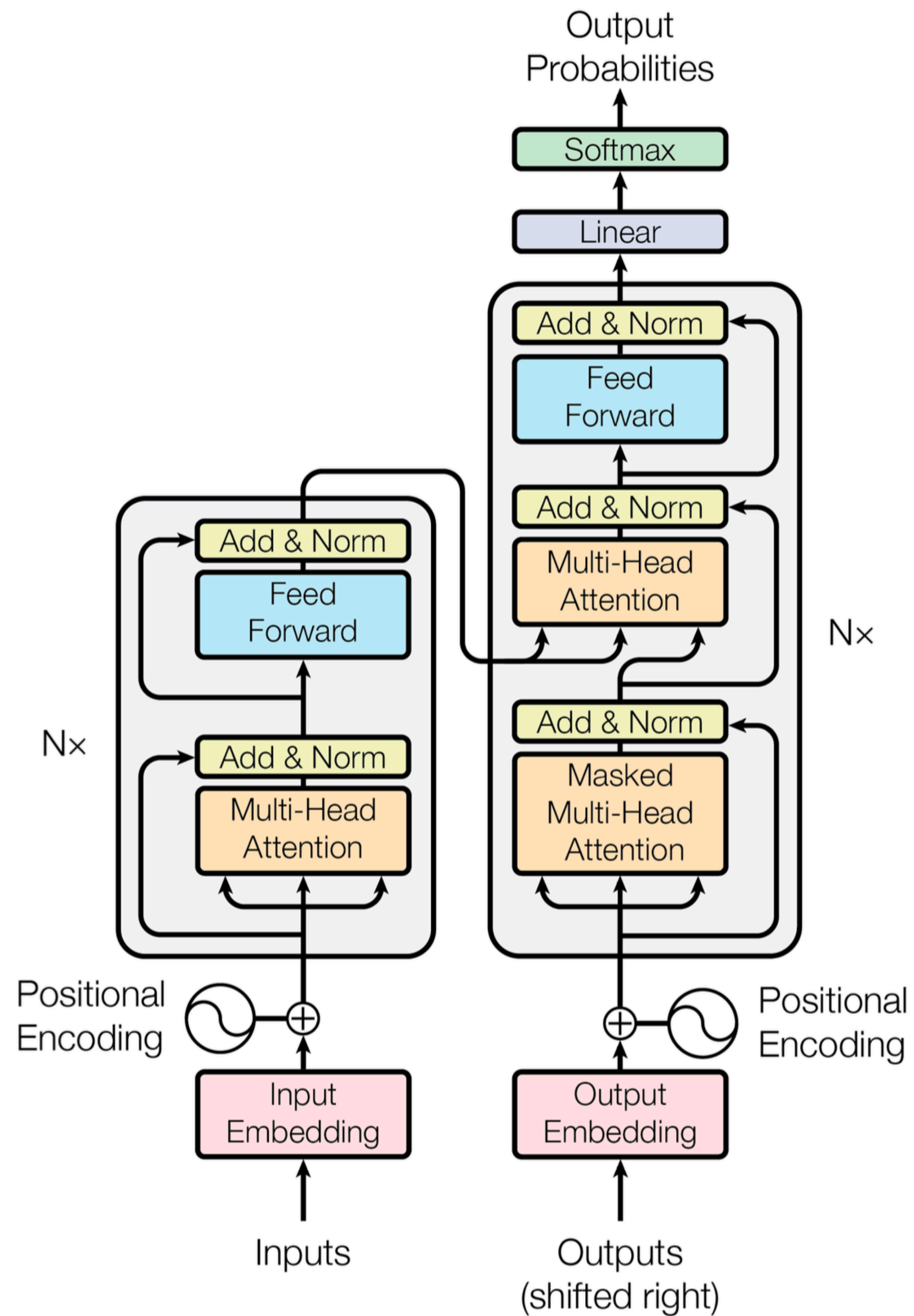
You can actually control LLMs

by activation steering

1. identify features — like honesty
2. during inference multiply them
3. model starts acting more honestly
4. to some extent

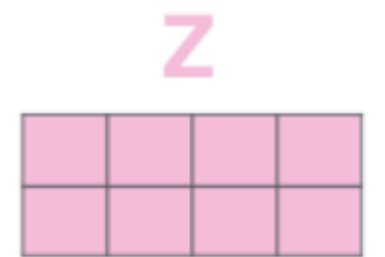
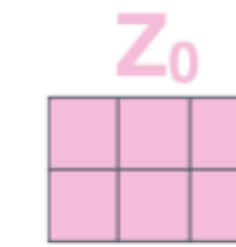
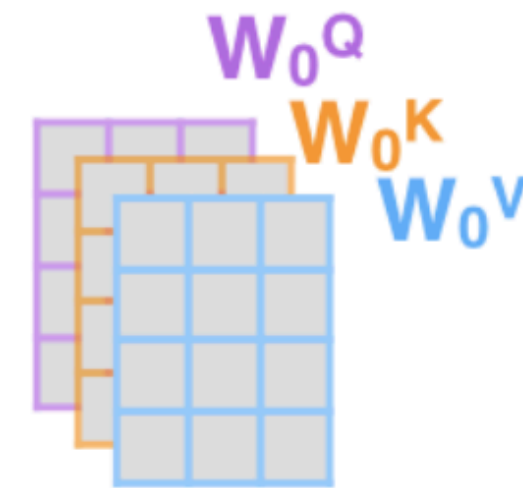
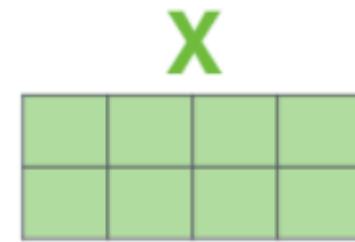
II. Circuits

Transformer recap

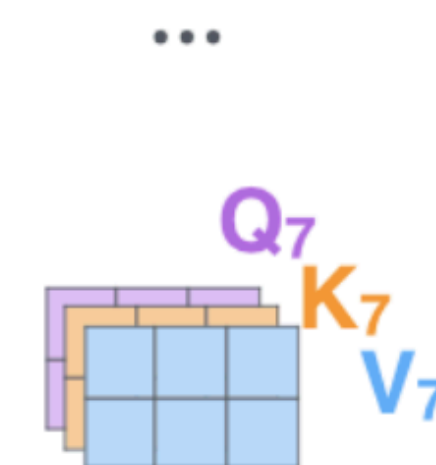
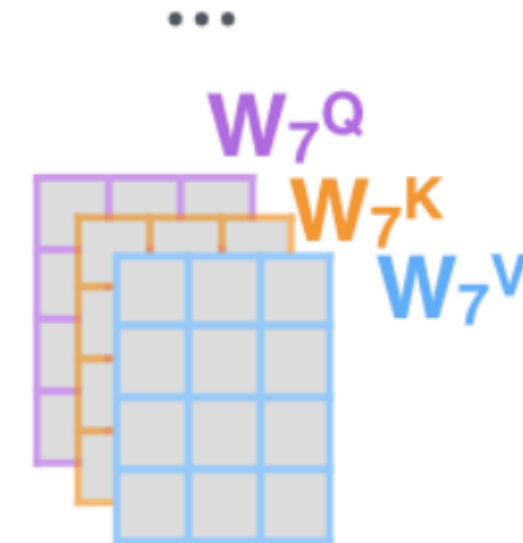
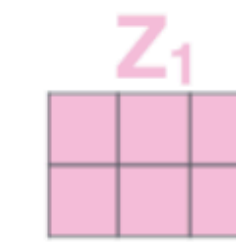
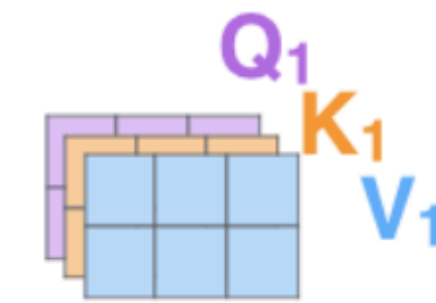
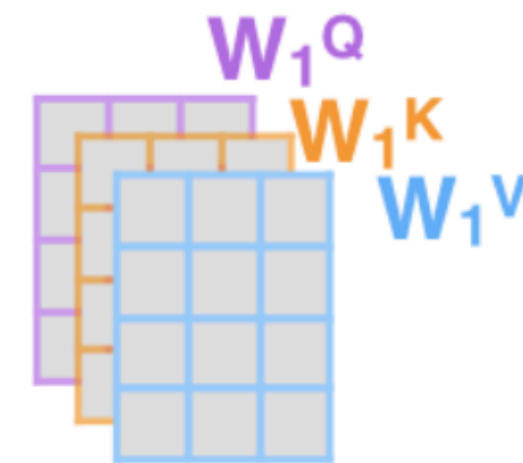


multi head atten

Thinking
Machines



* In all encoders other than #0,
we don't need embedding.
We start directly with the output
of the encoder right below this one

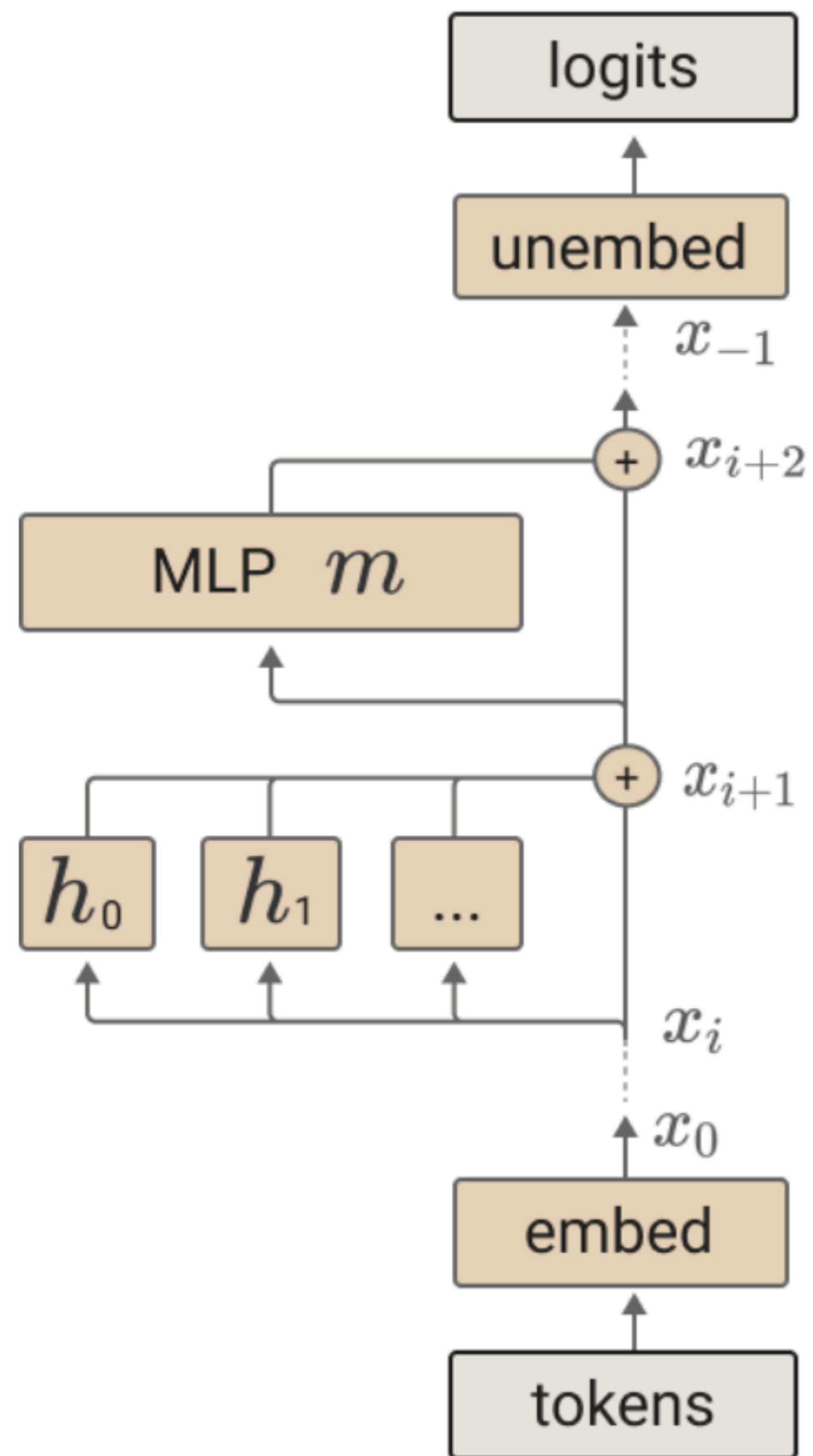


Remember?

Remember?

now forget :D

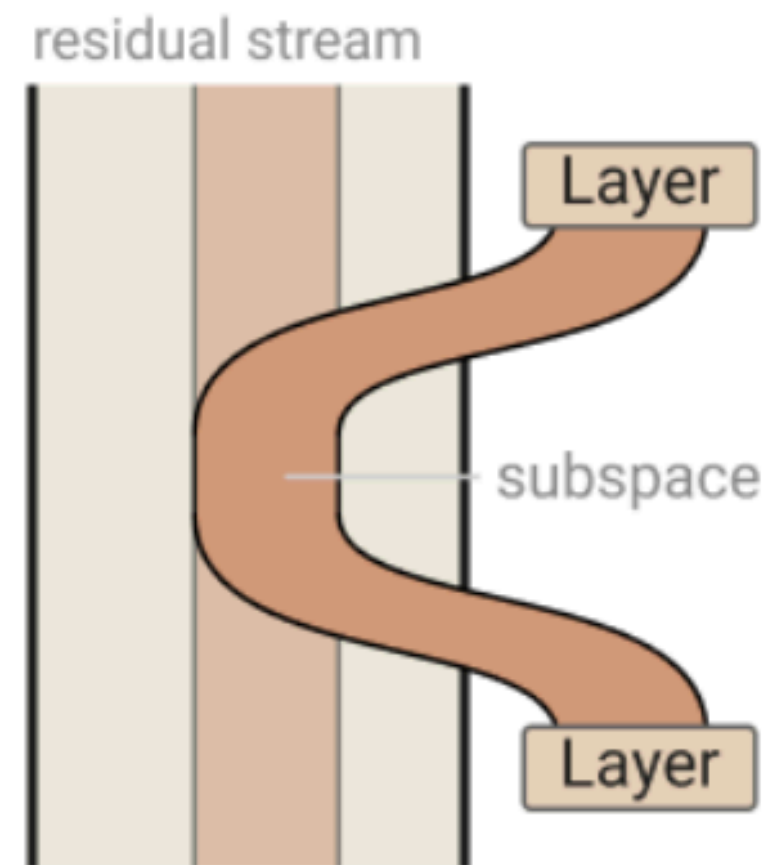
The right image
what changed?



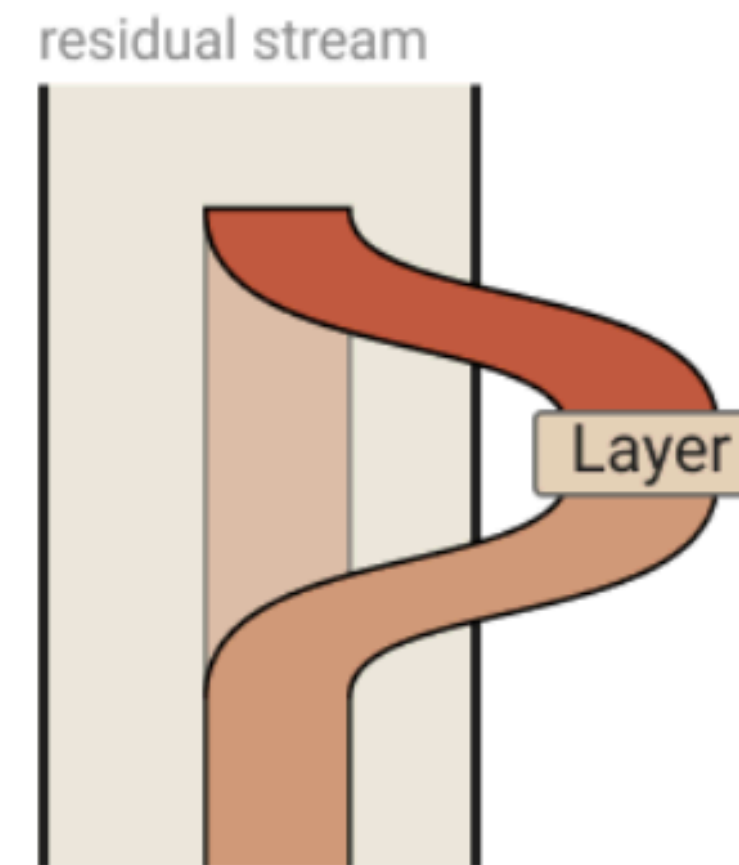
Residual stream

- Generally, it is an communication channel between layers
- W_k , W_q , W_v - read
- W_o - write

The residual stream is high dimensional, and can be divided into different subspaces.



Layers can interact by writing to and reading from the same or overlapping subspaces. If they write to and read from disjoint subspaces, they won't interact. Typically the spaces only partially overlap.



Layers can delete information from the residual stream by reading in a subspace and then writing the negative version.

Residual stream

has some structure

token encoding subspace (i.e. "this token is x ")

= rows of W_E

positional encoding subspace (i.e. "this token is at position x ")

= rows of W_{pos}

decoding subspace (i.e. "the next token will be x ")

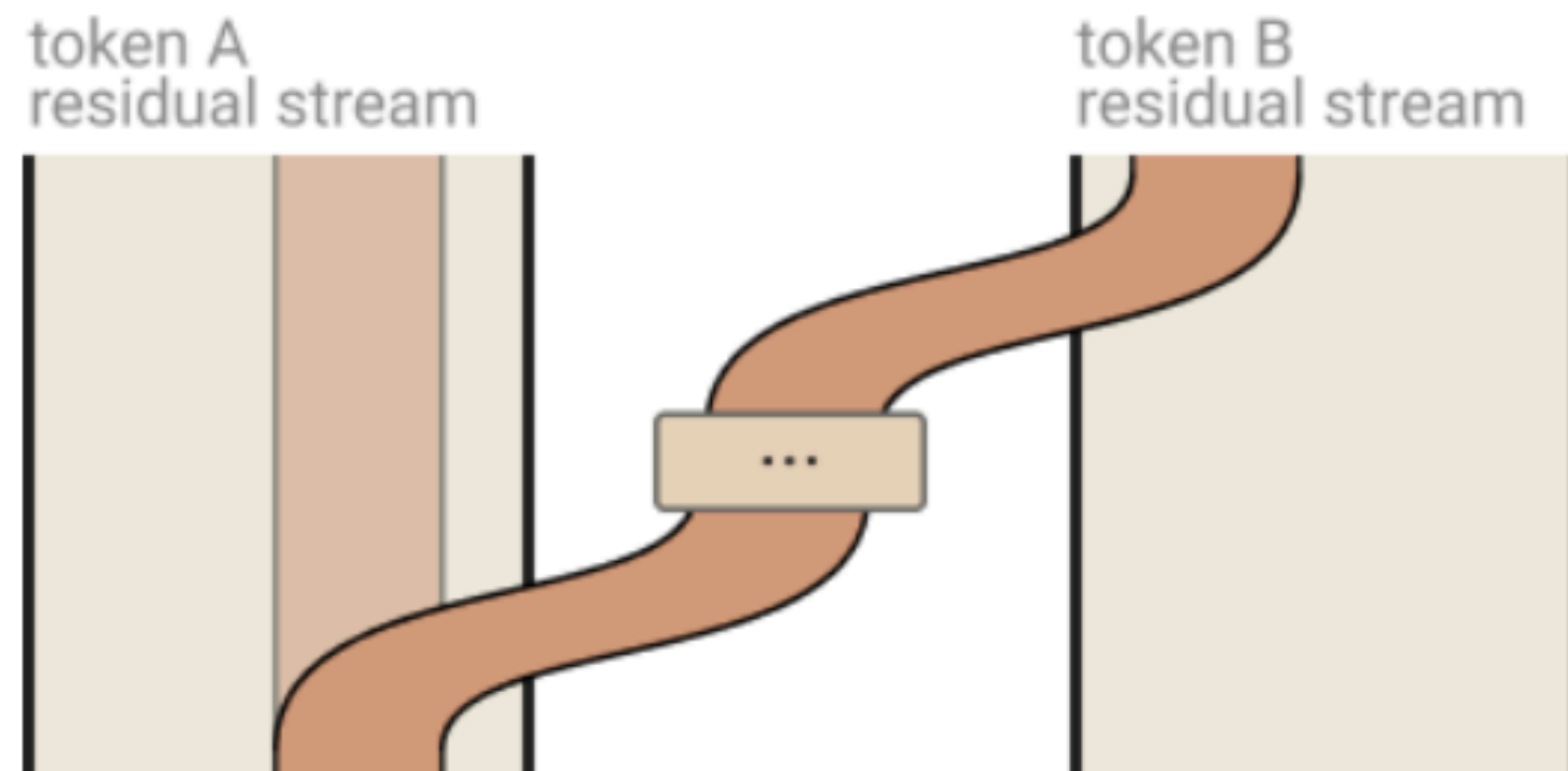
= cols of W_U

prev token subspace (i.e. "the previous token was x ")

= "intermediate information"

attention as information movement

attention as information movement



Attention heads copy information from the residual stream of one token to the residual stream of another. They typically write to a different subspace than they read from.

QK, VO circuits

MHA recap

$x, y \in \mathbb{R}^{d_{model}}$ - embeddings

$W_Q, W_K, W_V, W_O \in \mathbb{R}^{d_{head} \times d_{model}}$ — query, keys, values and output matrixes

$W_Q x, W_K x, W_V x, W_O x \in \mathbb{R}^{d_{head}}$ — query, keys, values and output vectors

QK, VO circuits

MHA recap

$x, y \in \mathbb{R}^{d_{model}}$ - embeddings

$W_Q, W_K, W_V, W_O \in \mathbb{R}^{d_{head} \times d_{model}}$ — query, keys, values and output matrixes

$W_Q x, W_K x, W_V x, W_O x \in \mathbb{R}^{d_{head}}$ — query, keys, values and output vectors

then we calculate the attention scores like this: $f_{score}(x, y) = y^T W_Q^T W_K x$

QK, VO circuits

MHA recap

$x, y \in \mathbb{R}^{d_{model}}$ - embeddings

$W_Q, W_K, W_V, W_O \in \mathbb{R}^{d_{head} \times d_{model}}$ — query, keys, values and output matrixes

$W_Q x, W_K x, W_V x, W_O x \in \mathbb{R}^{d_{head}}$ — query, keys, values and output vectors

then we calculate the attention scores like this: $f_{score}(x, y) = y^T W_Q^T W_K x$

we can define new matrix $W_{QK} := W_Q^T W_K$ and call it an «QK-circuit»

QK matrix is basically a bilinear form on embeddings

bilinear form

- $B(x, y) = xBy^T$, where $x, y \in \mathbb{R}^n$, $B \in \mathbb{R}^{n \times n}$
- $B(x, y) : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$

QK, VO circuits

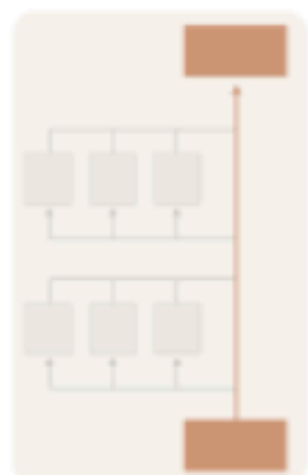
are bilinear forms!

- words are vectors
- QK - how much information to move from x to y
- VO - what information to move from x to y

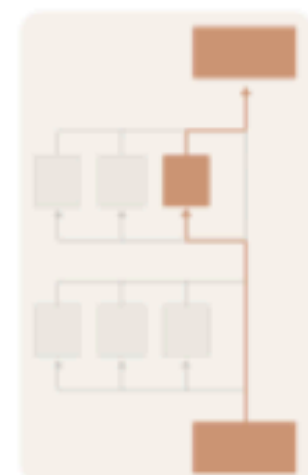
heads form some circuits

what?

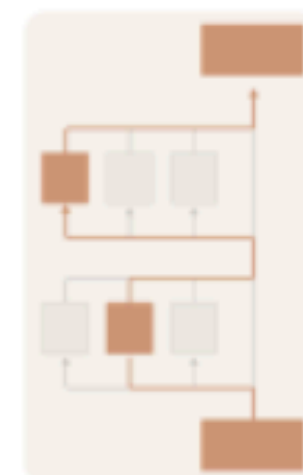
$$= \underbrace{\text{Id} \otimes W_U W_E}_{\text{Diagram 1}} + \sum_{h \in H_1 \cup H_2} A^h \otimes (W_U W_{OV}^h W_E)_{\text{Diagram 2}} + \sum_{h_2 \in H_2} \sum_{h_1 \in H_1} (A^{h_2} A^{h_1}) \otimes (W_U W_{OV}^{h_2} W_{OV}^{h_1} W_E)_{\text{Diagram 3}}$$



“Direct path” term contributes to bigram statistics.



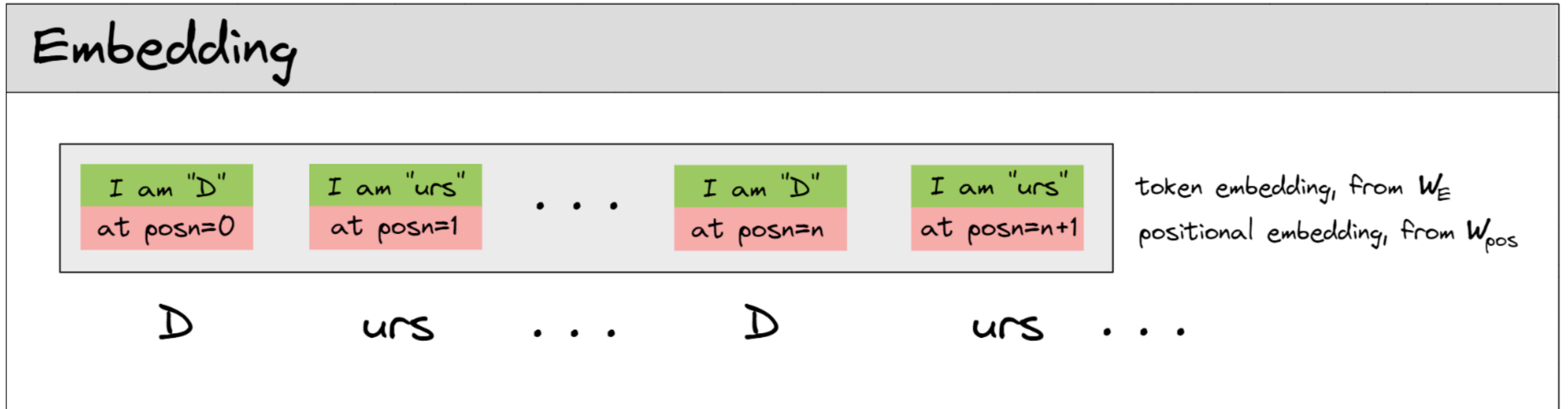
The **individual attention head** terms describe the effects of individual attention heads in linking input tokens to logits, similar to those we saw in the one layer model.



The **virtual attention head** terms correspond to V-composition of attention heads. They function a lot like individual attention heads, with their own attention patterns (the composition of the heads patterns) and own OV matrix.

III. Circuits Examples

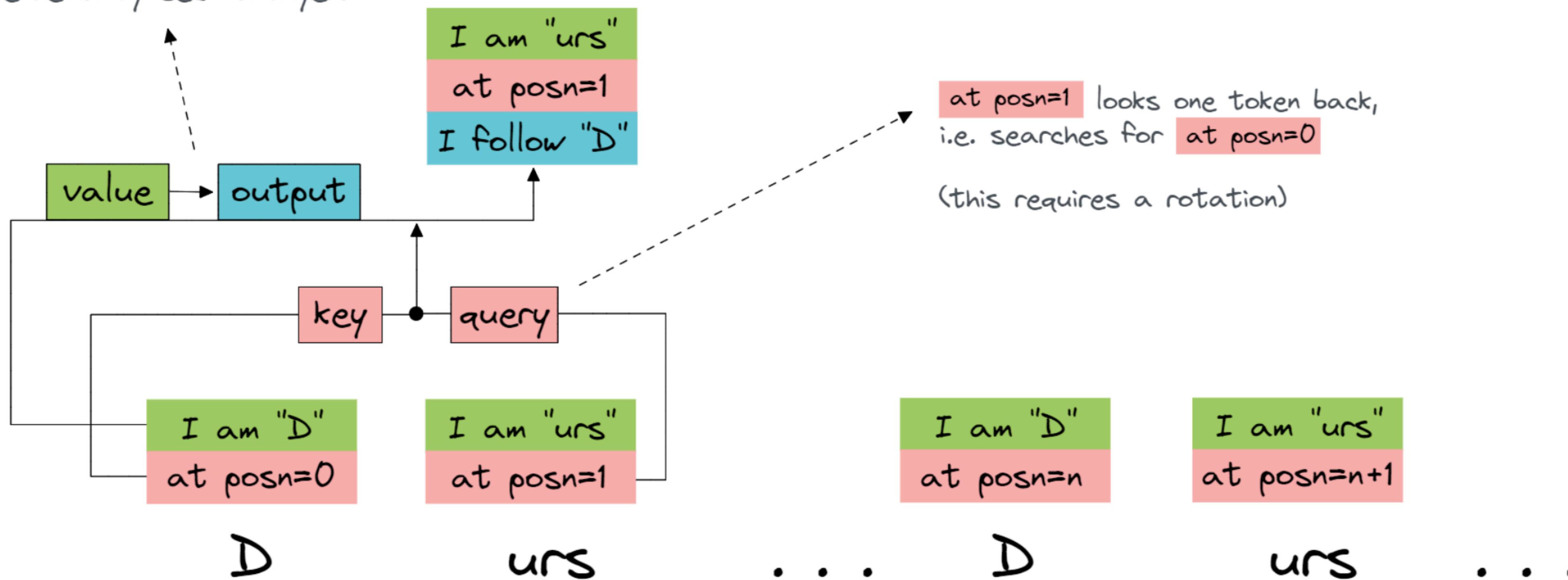
induction heads



induction heads

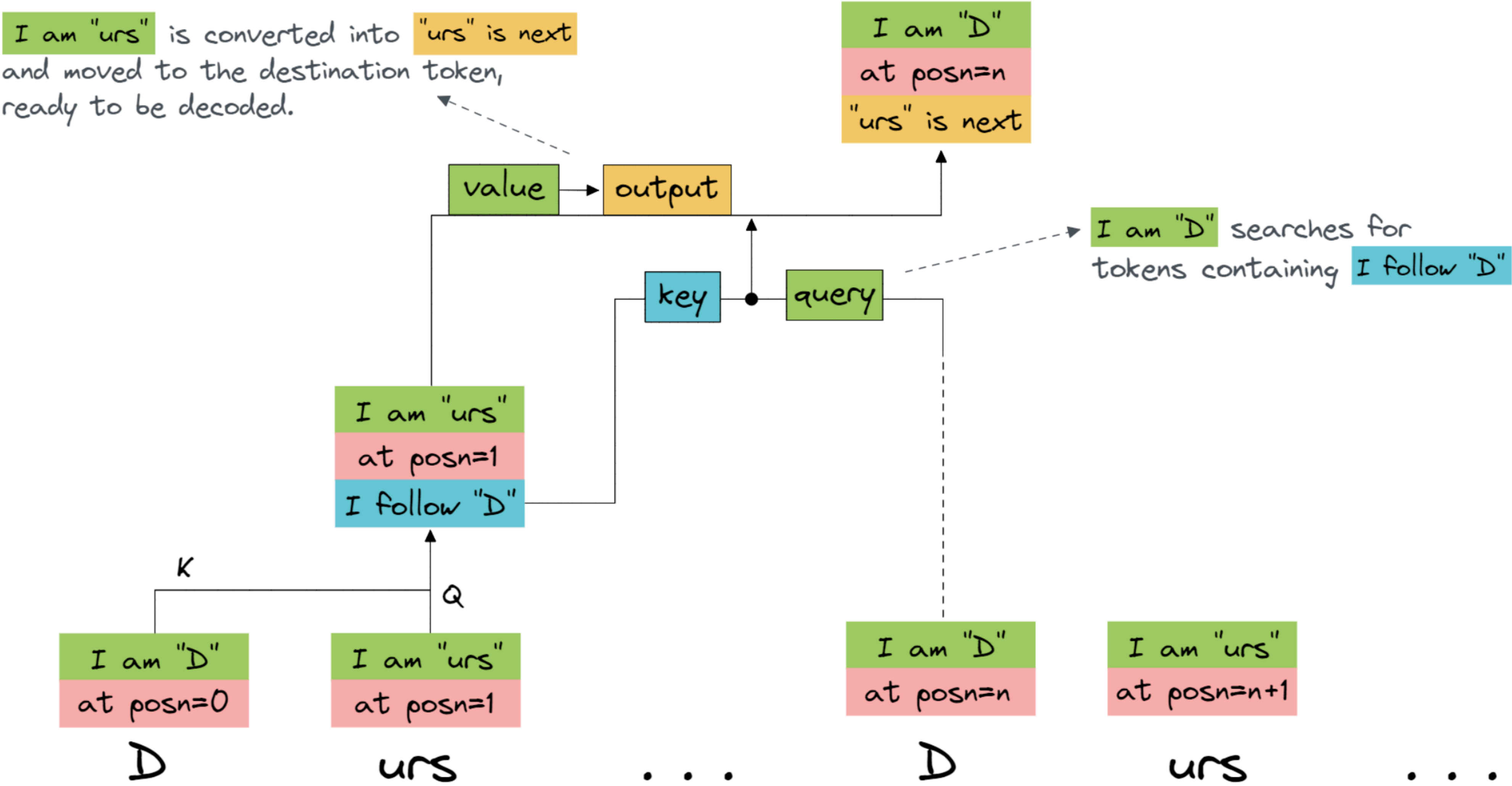
layer 0

I am "D" is converted to **I follow "D"**
and moved to the destination token,
ready to be read by second layer.



induction heads

layer 1



Some circuits examples

modular addition

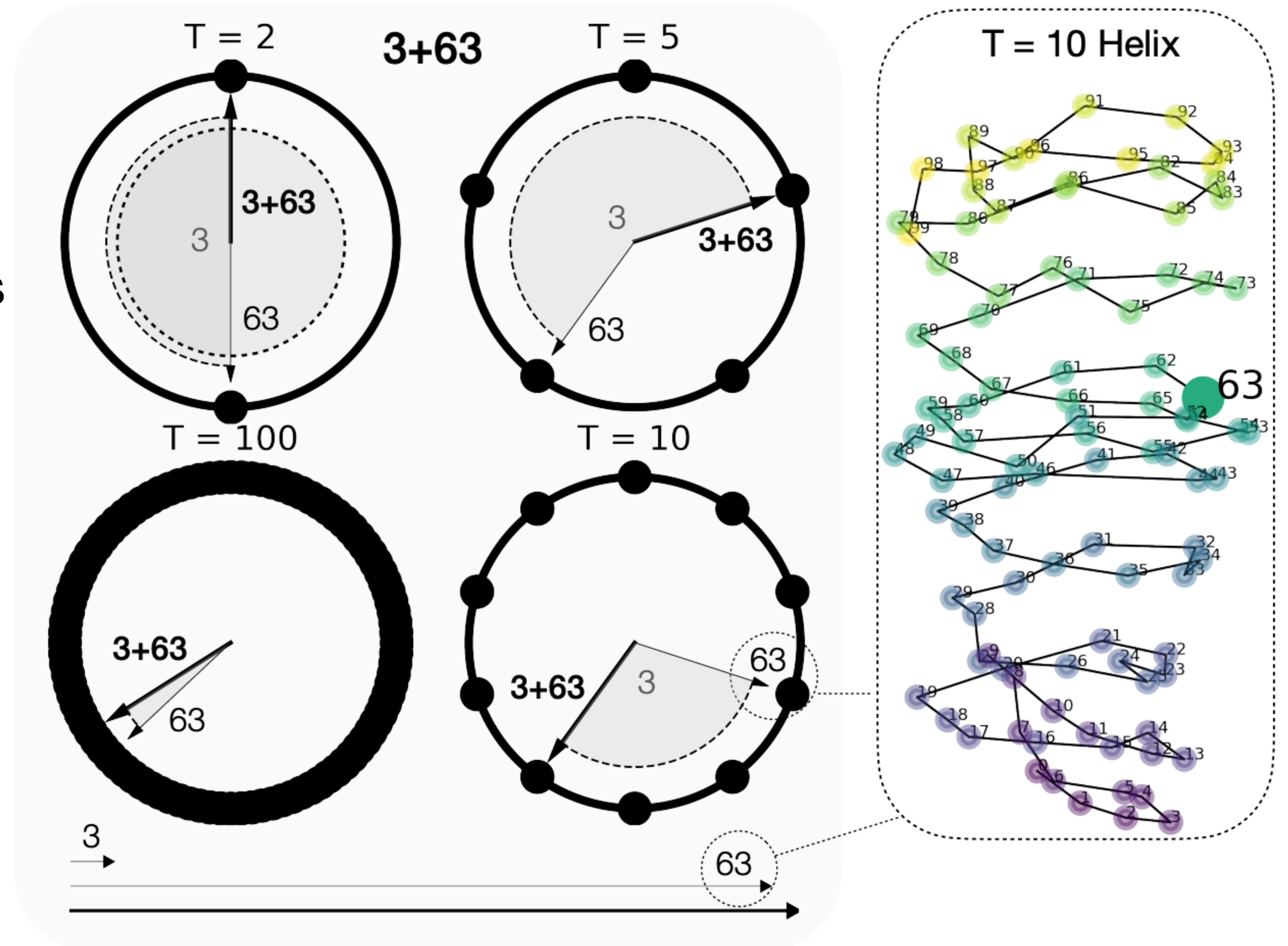
$$a + b =$$

1. embed a and b on circles

[2,5,10,100]

2. sum them on circles

3. read back to logits



Some circuits examples

- indirect object identification (IOI)
- fact localisation

IV. Thanks for your attention :)

materials

- https://transformer-circuits.pub/2022/toy_model/index.html
- <https://www.lesswrong.com/posts/TvrfY4c9eaGLeyDkE/induction-heads-illustrated>
- <https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html>
- <https://arxiv.org/abs/2502.00873>
- <https://transformer-circuits.pub/2021/framework/index.html>