

Mechanistic Permutability

Nikita Balagansky, Ian Maksimov, Daniil Gavrilov

Interpreting Residual Stream of the LLMs

Non-privileged basis

There is no any activations in the residual stream.

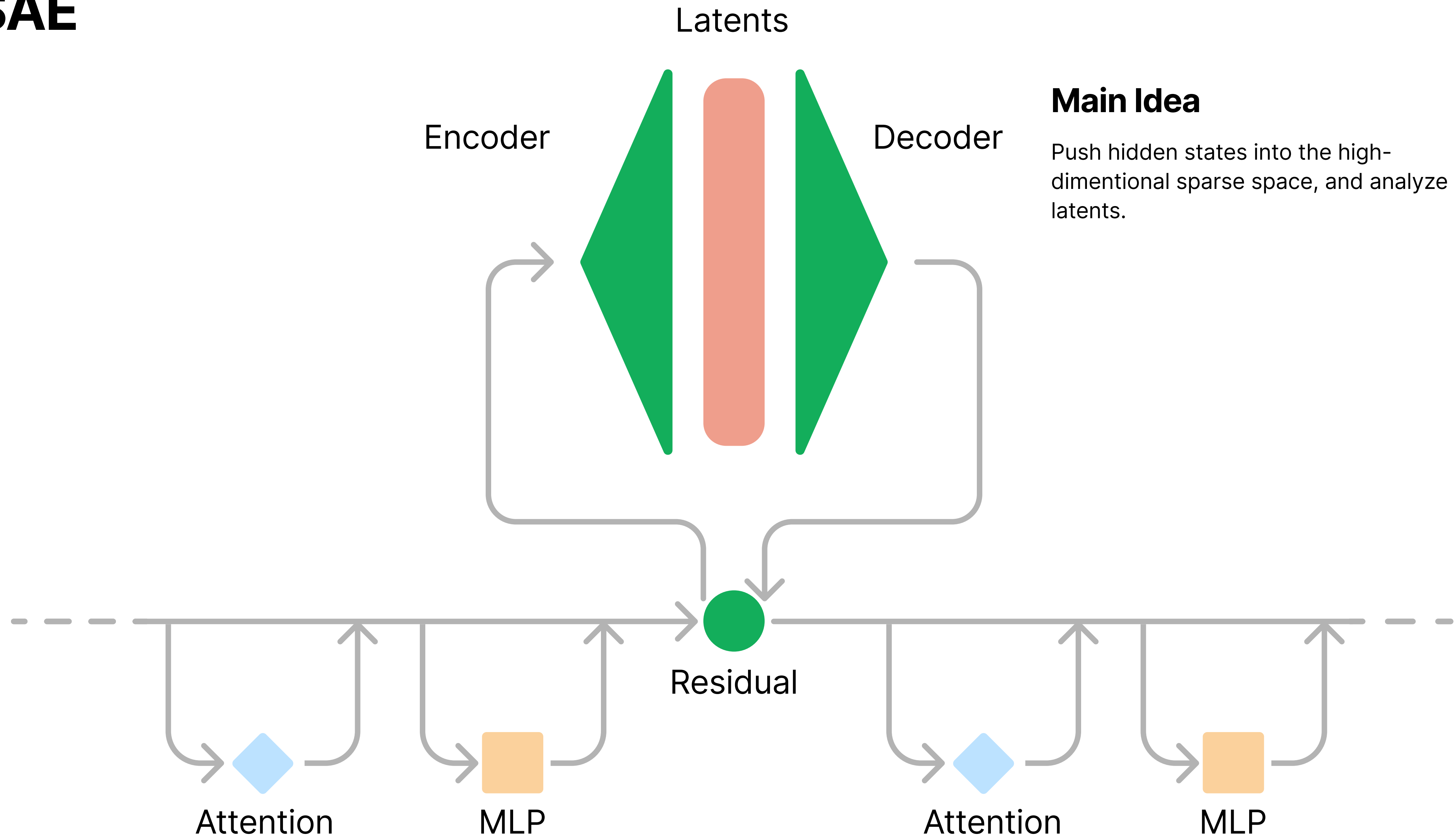
→ Polysemanticity

A single “neuron” fires in seemingly random locations across the text.

Features superposition

Features on language model exists, but it could be depend on others.

SAE



Main Idea

Push hidden states into the high-dimensional sparse space, and analyze latents.

SAE

Latents

Encoder

Decoder

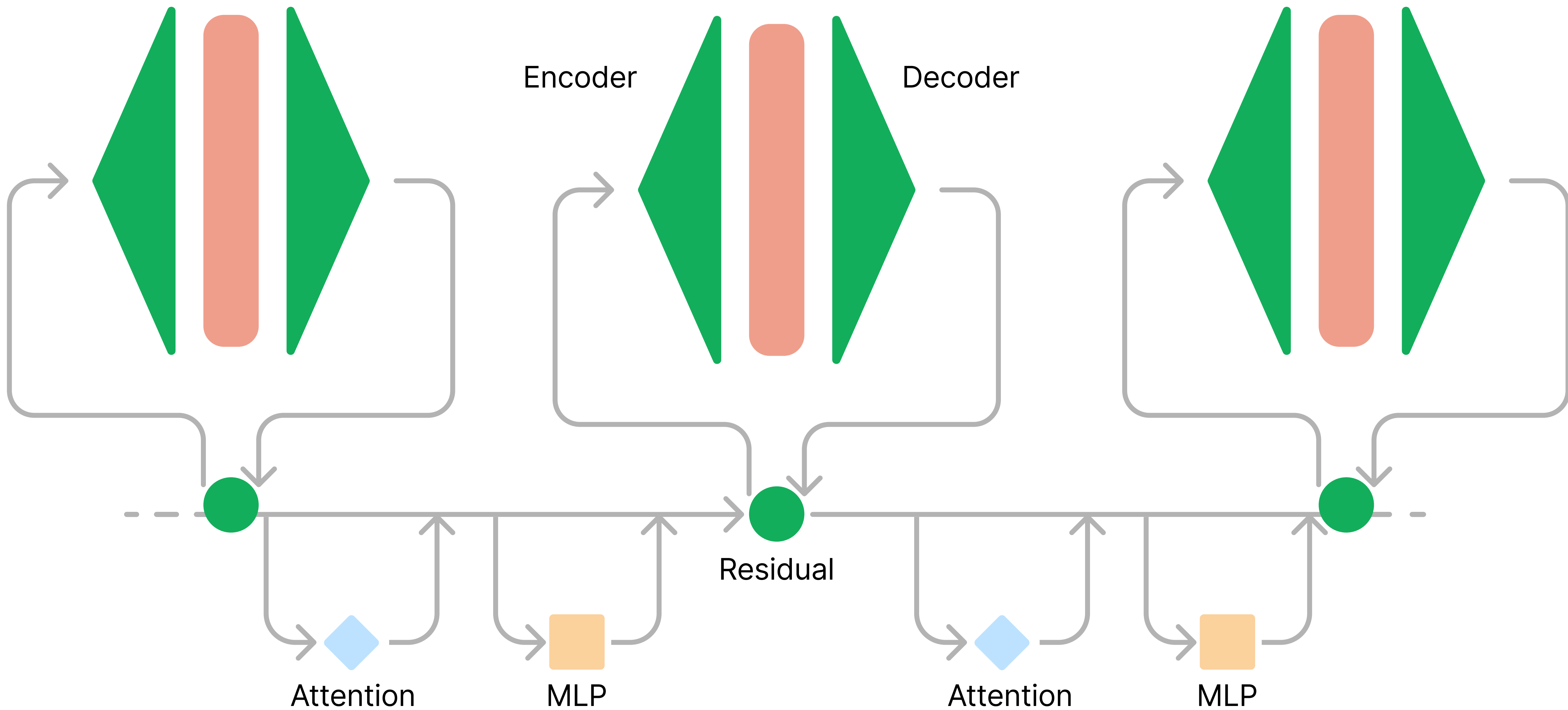
Residual

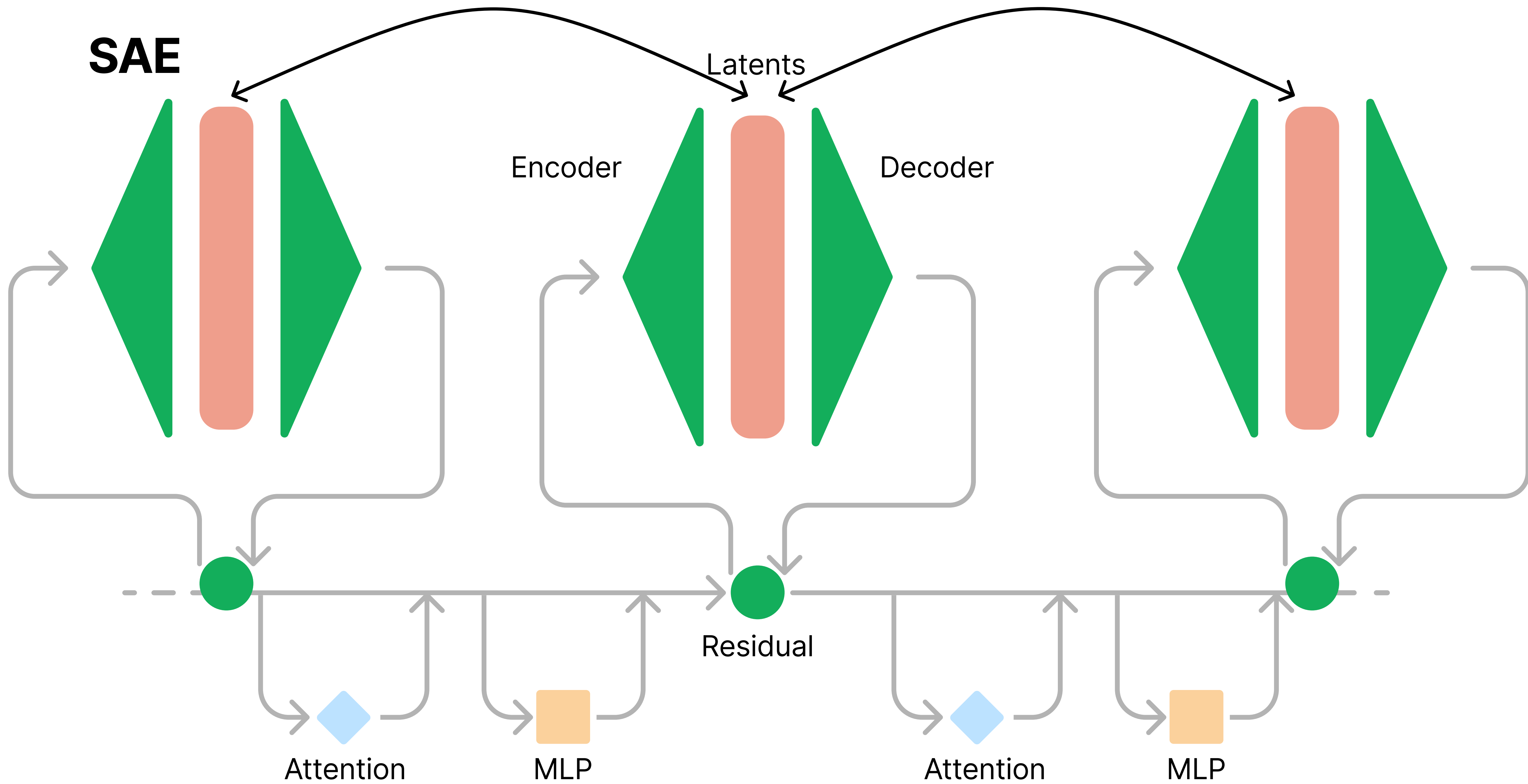
Attention

MLP

Attention

MLP

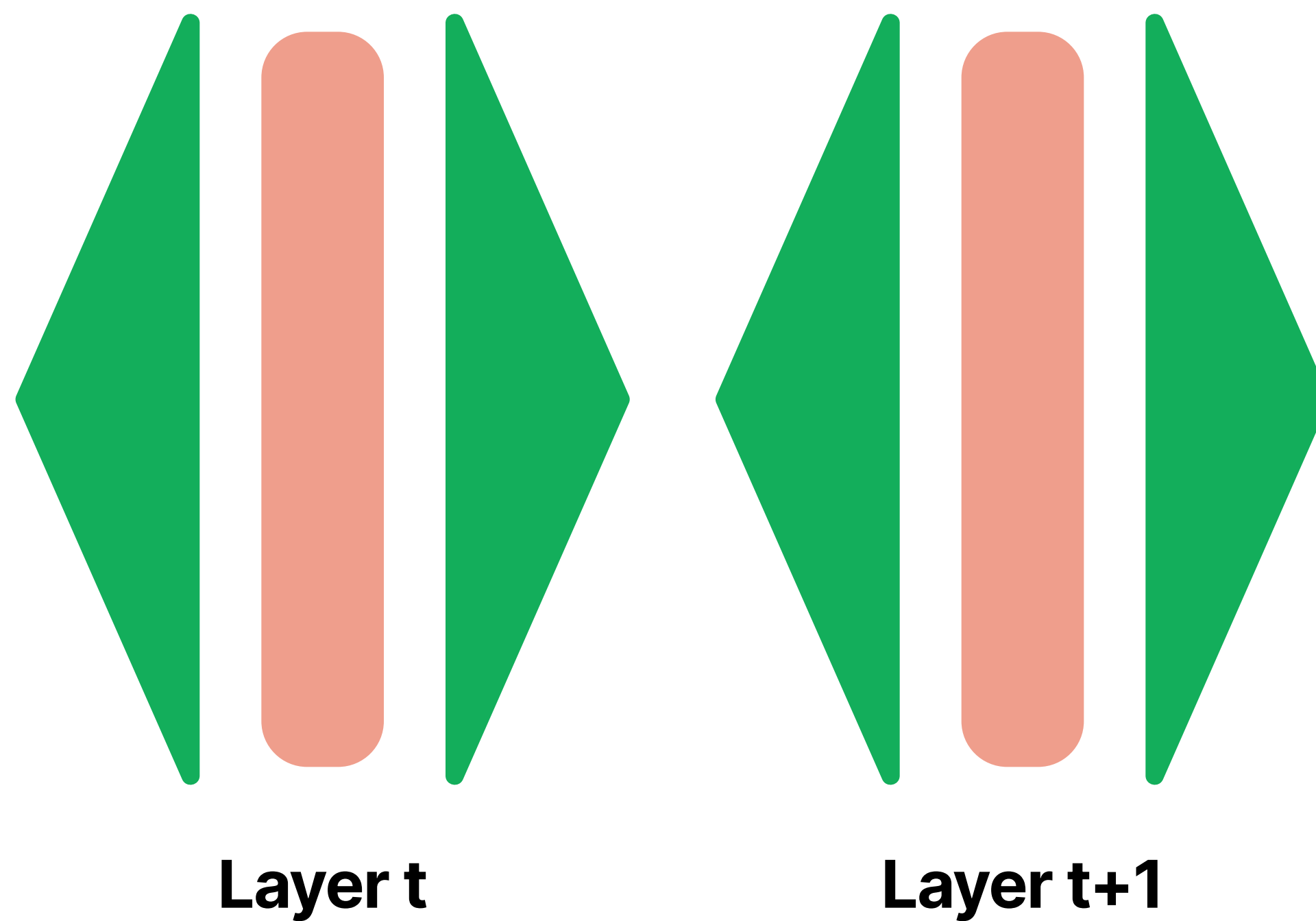




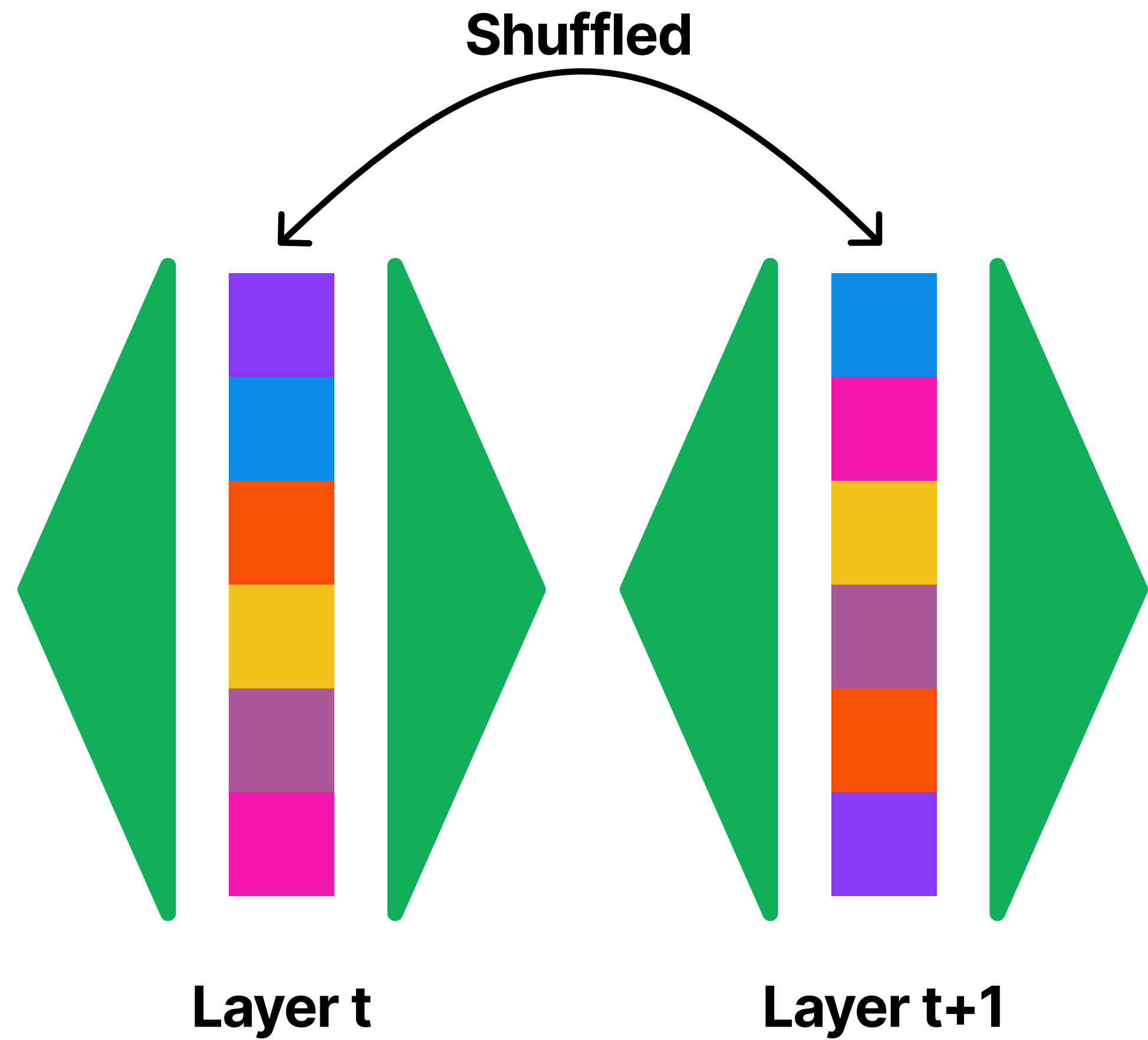
Hypothesis

- Most of the features remain the same in the residual stream during forward pass.
- Same features on the different layers have the same embedding.

**There is a problem with
latents on different
layers.**

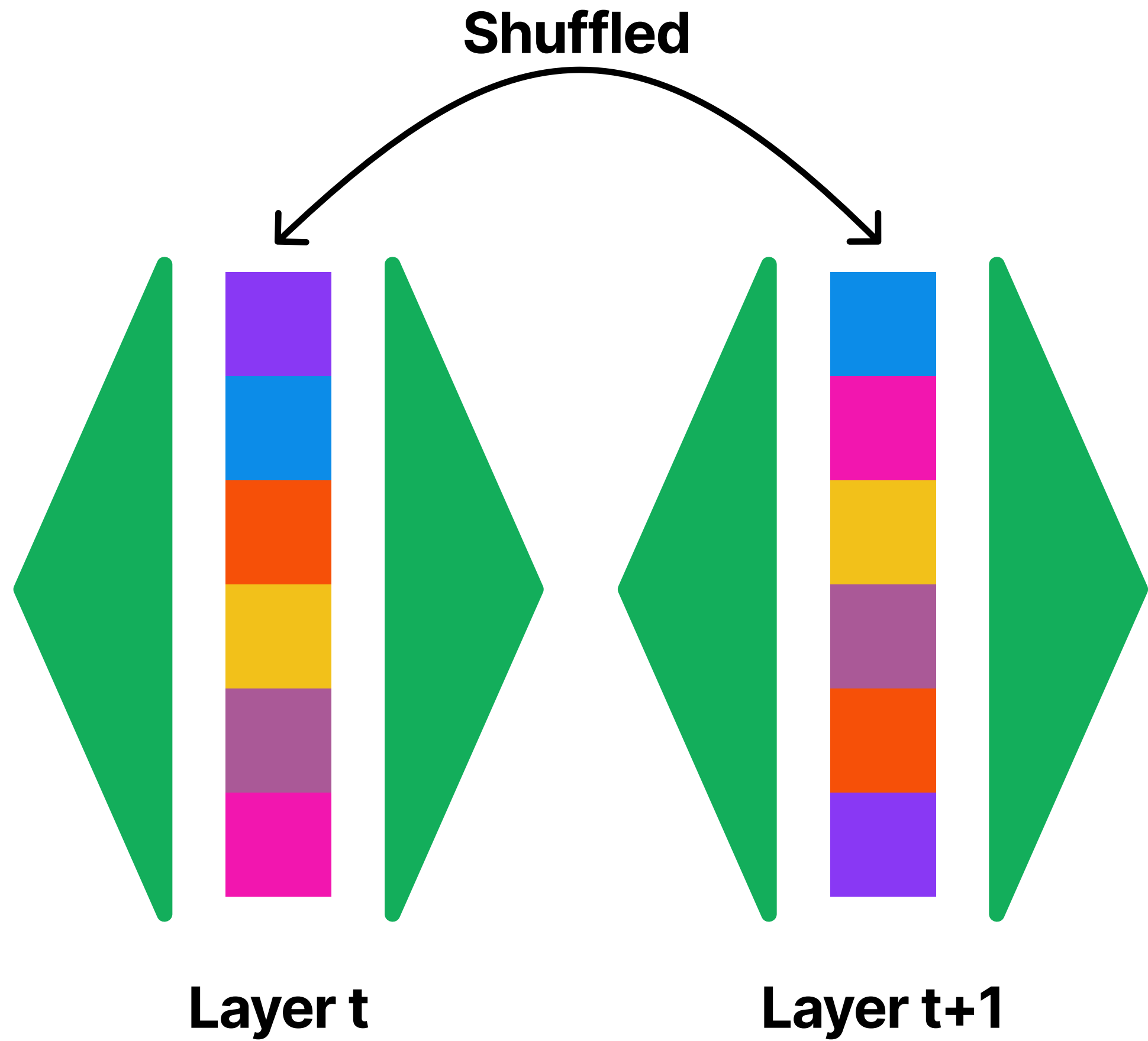


**There is a problem with
latents on different
layers.**



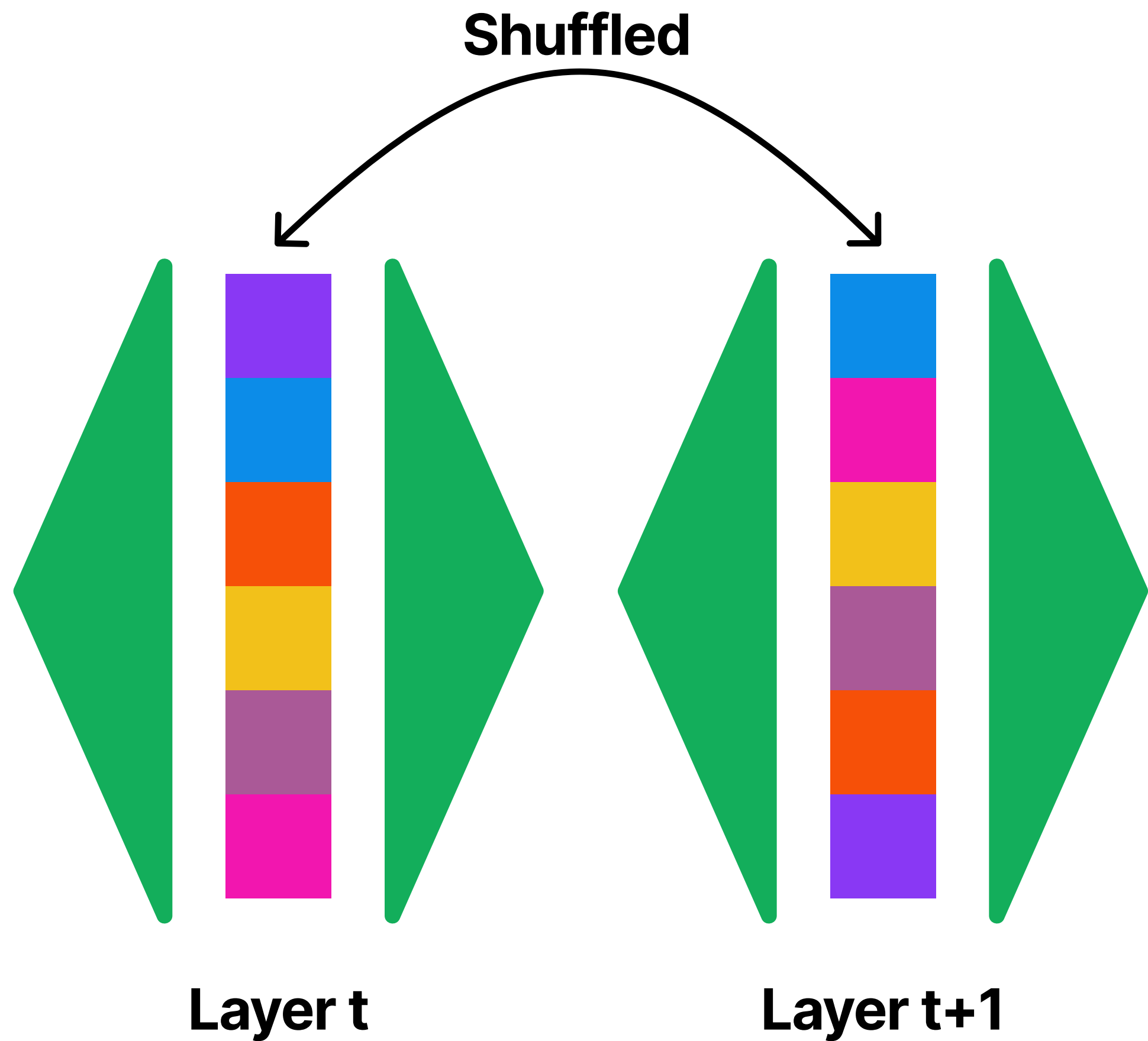
There is a problem with latents on different layers.

- Find correlation between features



There is a problem with latents on different layers.

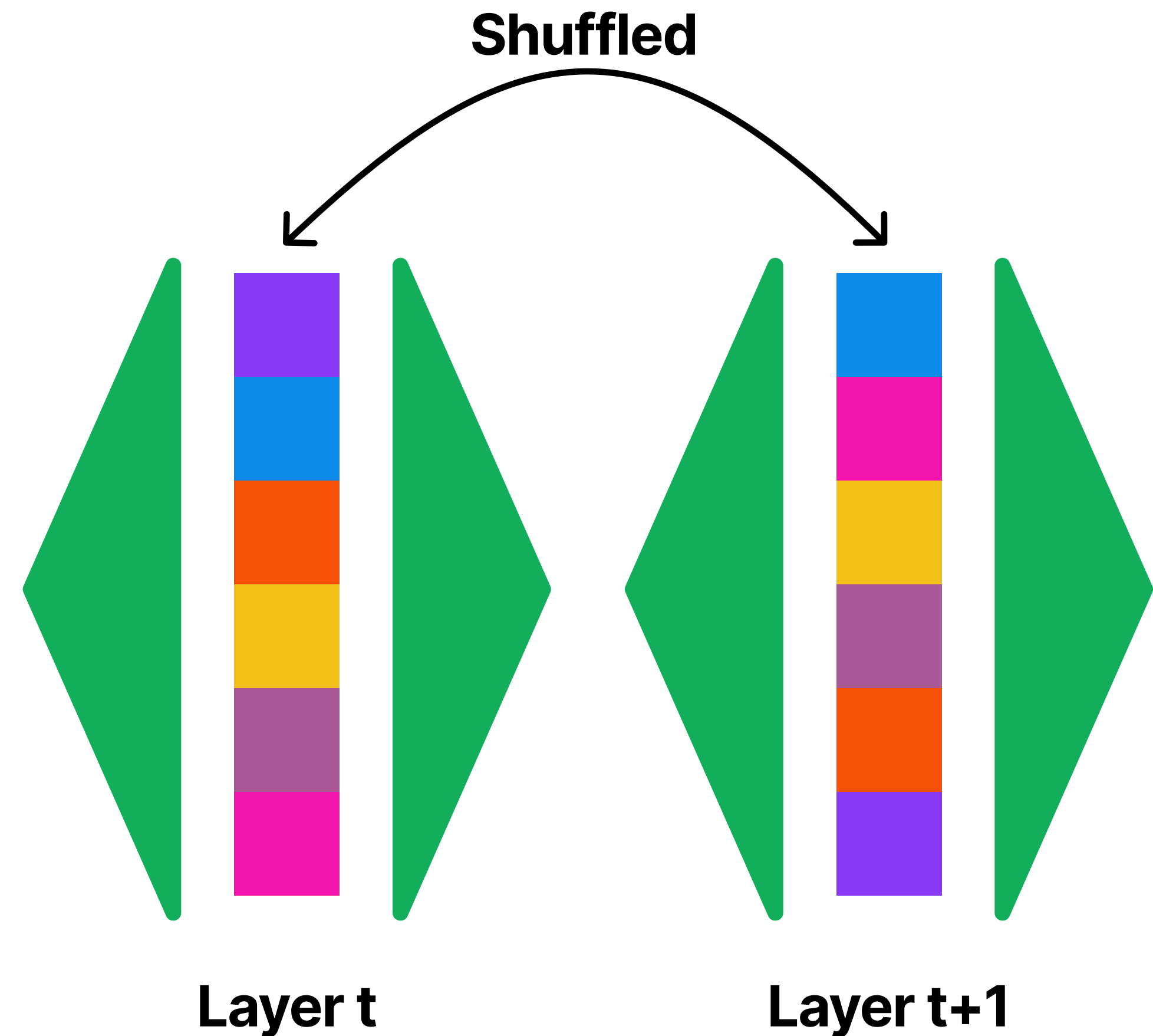
- Find correlation between features
(requires data and computational resources)
- Use weights to find permutation



There is a problem with latents on different layers.

- Find correlation between features
(requires data and computational resources)
- Use weights to find permutation

$$\arg \min_{P \in \mathcal{P}_F} \sum_{i=1}^d \left\| \mathbf{W}_{\text{dec}_{i,:}}^{(A)} - P \mathbf{W}_{\text{dec}_{i,:}}^{(B)} \right\|^2$$



Obtained example

- Find correlation between features
(requires data and computational resources)
- Use weights to find permutation

$$\arg \min_{P \in \mathcal{P}_F} \sum_{i=1}^d \left\| \mathbf{W}_{\text{dec}_{i,:}}^{(A)} - P \mathbf{W}_{\text{dec}_{i,:}}^{(B)} \right\|^2$$

time intervals.

GEMMA-2-2B
20-GEMMASCOPE-RES-16K
INDEX 4

NEGATIVE LOGITS ?

SBATCH

-0.59

POSITIVE LOGITS ?

time

ACT DENSITY 0.110% ?

1.03

1500

1000

phrases related to time durations and measurements

GEMMA-2-2B
21-GEMMASCOPE-RES-16K
INDEX 5748

NEGATIVE LOGITS ?

Normdatei

-0.81

POSITIVE LOGITS ?

maximum

ACT DENSITY 0.381% ?

1.68

6000

4000

phrases related to time duration and scheduling

GEMMA-2-2B
22-GEMMASCOPE-RES-16K
INDEX 10226

NEGATIVE LOGITS ?

Normdatei

-0.81

POSITIVE LOGITS ?

max

ACT DENSITY 0.310% ?

1.51

4000

3000

mentions of time durations and intervals

GEMMA-2-2B
23-GEMMASCOPE-RES-16K
INDEX 3065

NEGATIVE LOGITS ?

Normdatei

-0.64

POSITIVE LOGITS ?

max

ACT DENSITY 0.265% ?

1.36

3000

2000

quantitative measurements related to time and distance

GEMMA-2-2B
24-GEMMASCOPE-RES-16K
INDEX 16005

NEGATIVE LOGITS ?

ousands

-0.68

POSITIVE LOGITS ?

max

ACT DENSITY 0.206% ?

1.31

2000

numerical time-related references, particularly years and months

GEMMA-2-2B
25-GEMMASCOPE-RES-16K
INDEX 11582

NEGATIVE LOGITS ?

BrowserModule

-0.64

POSITIVE LOGITS ?

later

ACT DENSITY 0.145% ?

1.15

3000

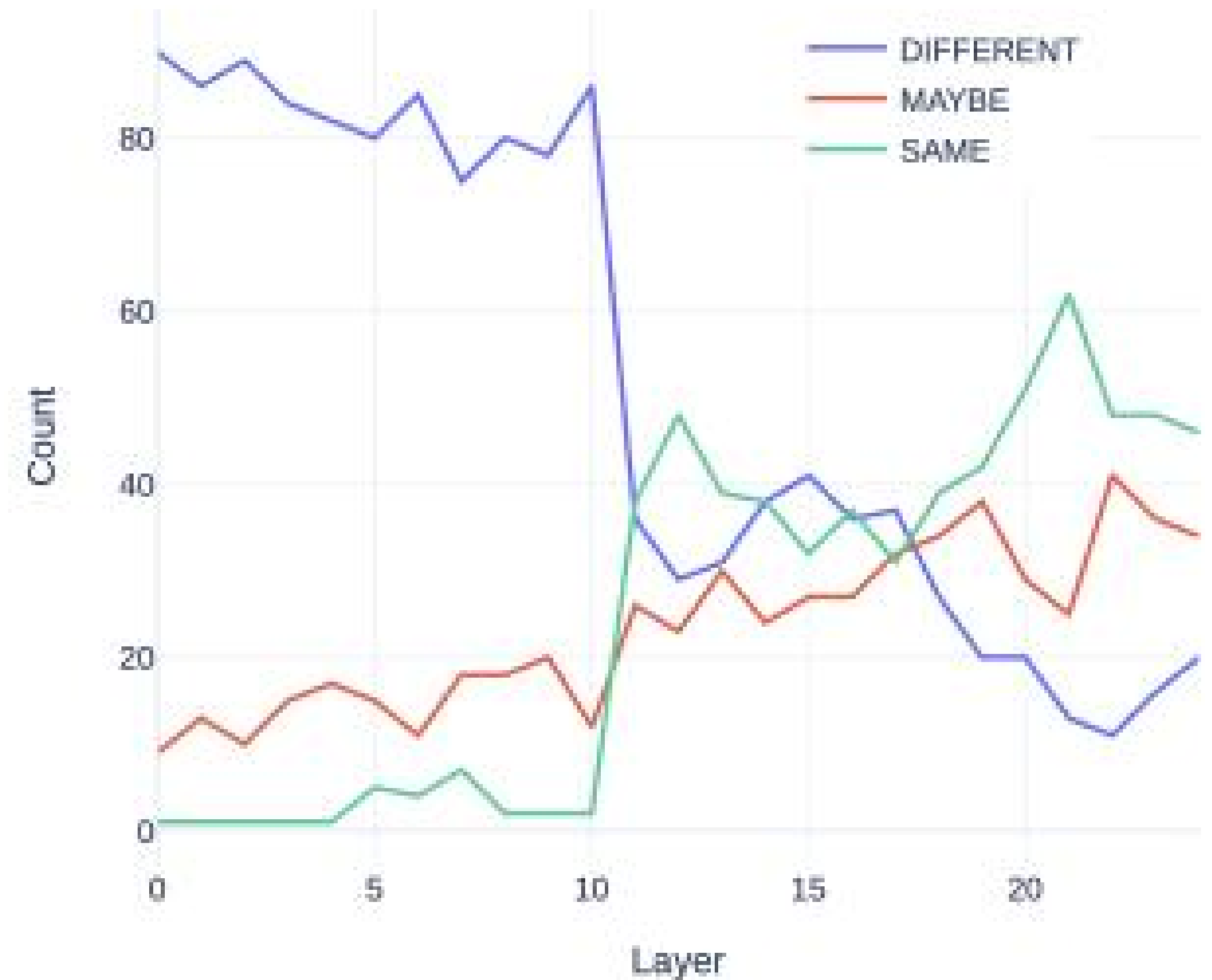
2000

Let's validate

- Find correlation between features
(requires data and computational resources)
- Use weights to find permutation

$$\arg \min_{P \in \mathcal{P}_F} \sum_{i=1}^d \left\| \mathbf{W}_{\text{dec}_{i,:}}^{(A)} - P \mathbf{W}_{\text{dec}_{i,:}}^{(B)} \right\|^2$$

Folded + Matched



**Check out the full
version here.**

