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.

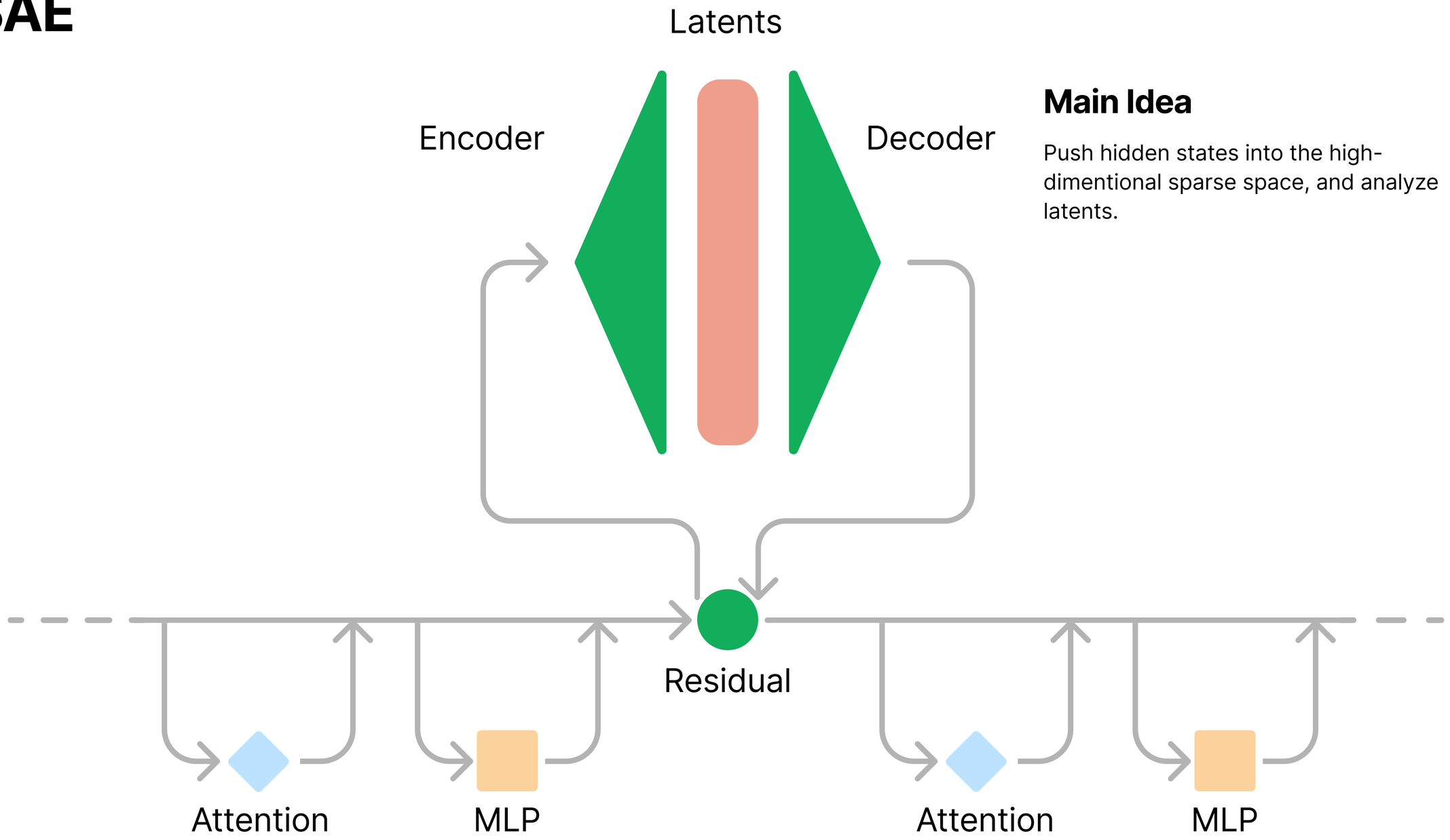
Polysemanticity

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

Features superposition

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

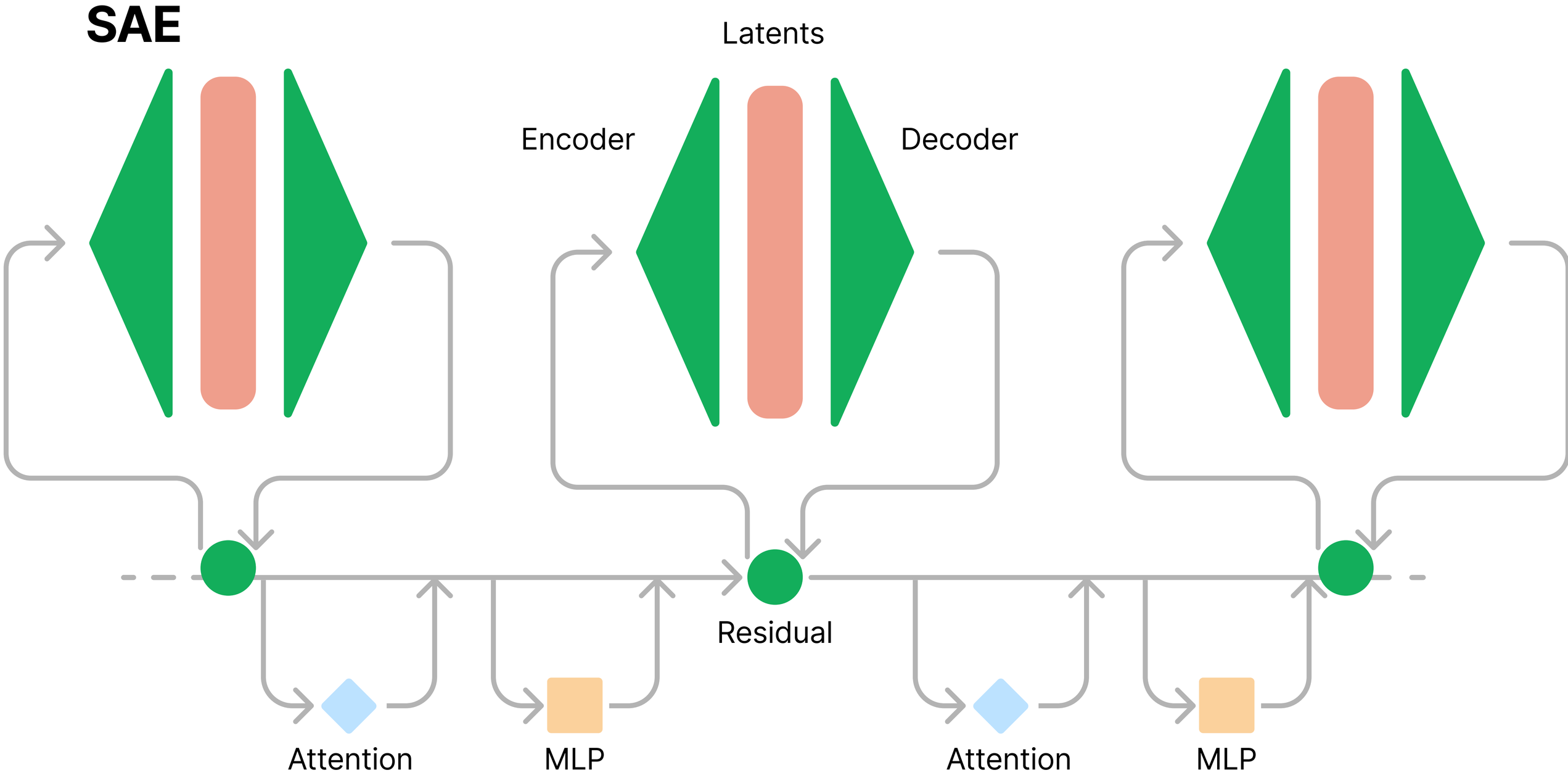
SAE

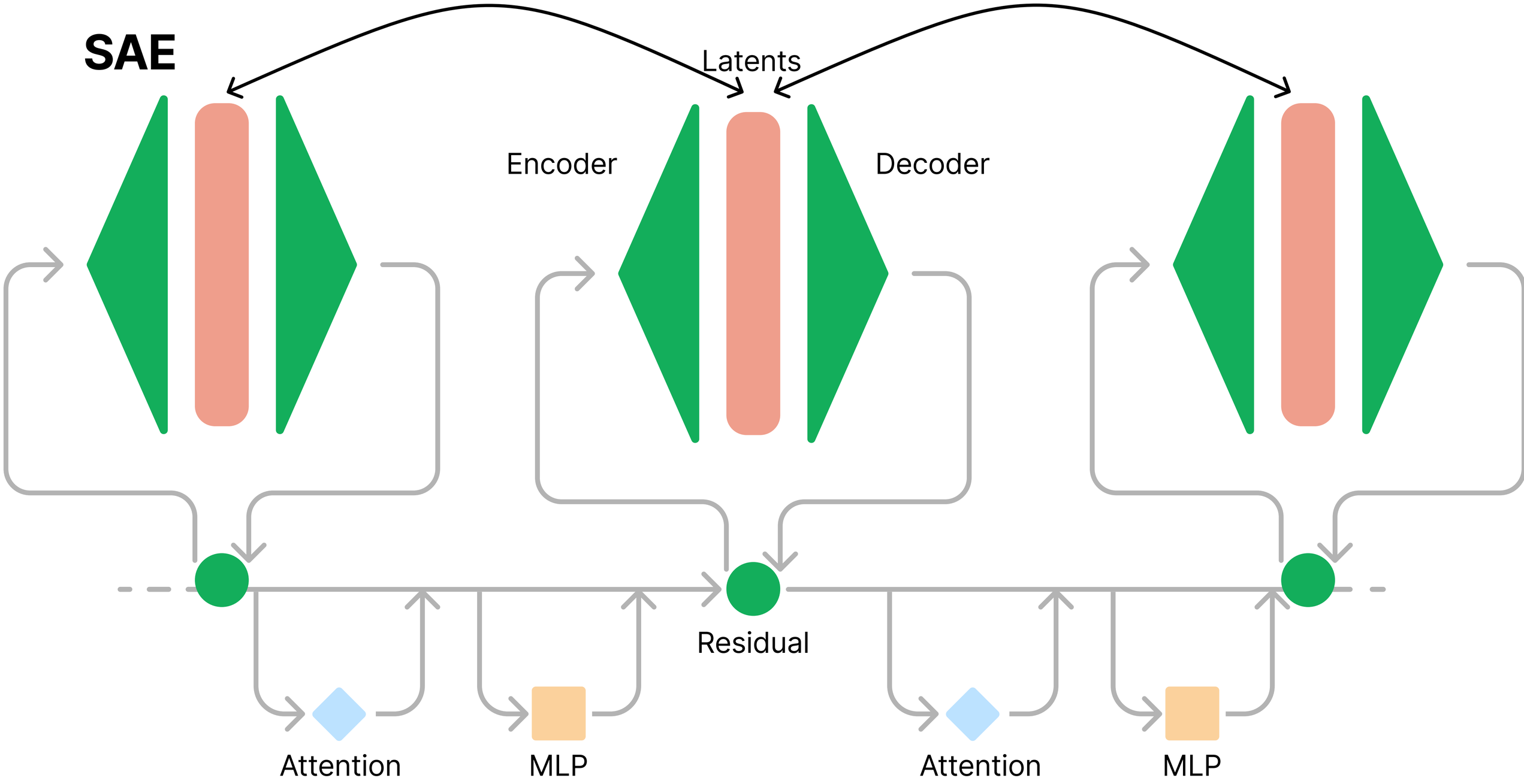


Main Idea

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

Adapted from https://www.goodfire.ai/papers/understanding-and-steering-llama-3

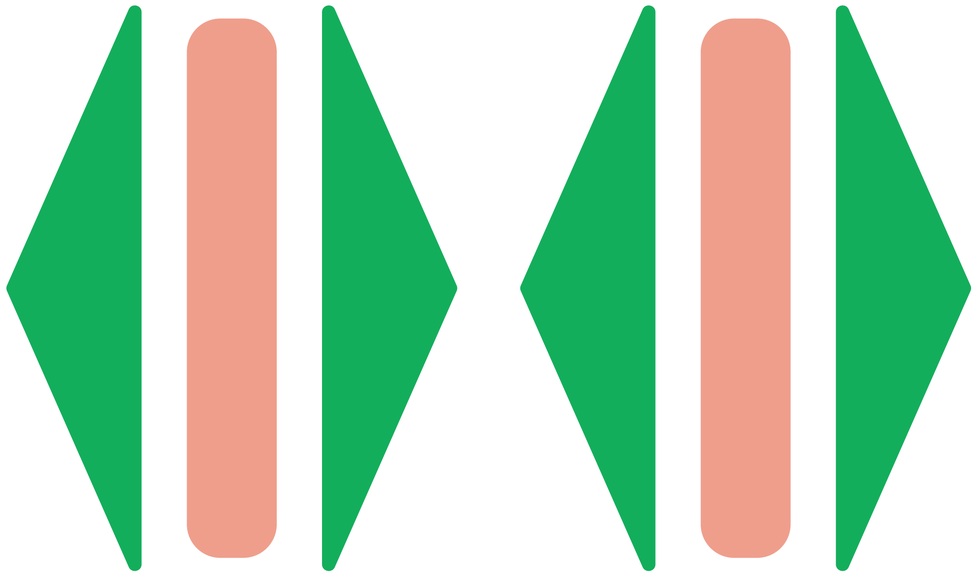

SAE
Latents
Encoder
Decoder
Residual
Attention
MLP
Attention
MLP


SAE
Latents
Encoder
Decoder
Residual
Attention
MLP
Attention
MLP

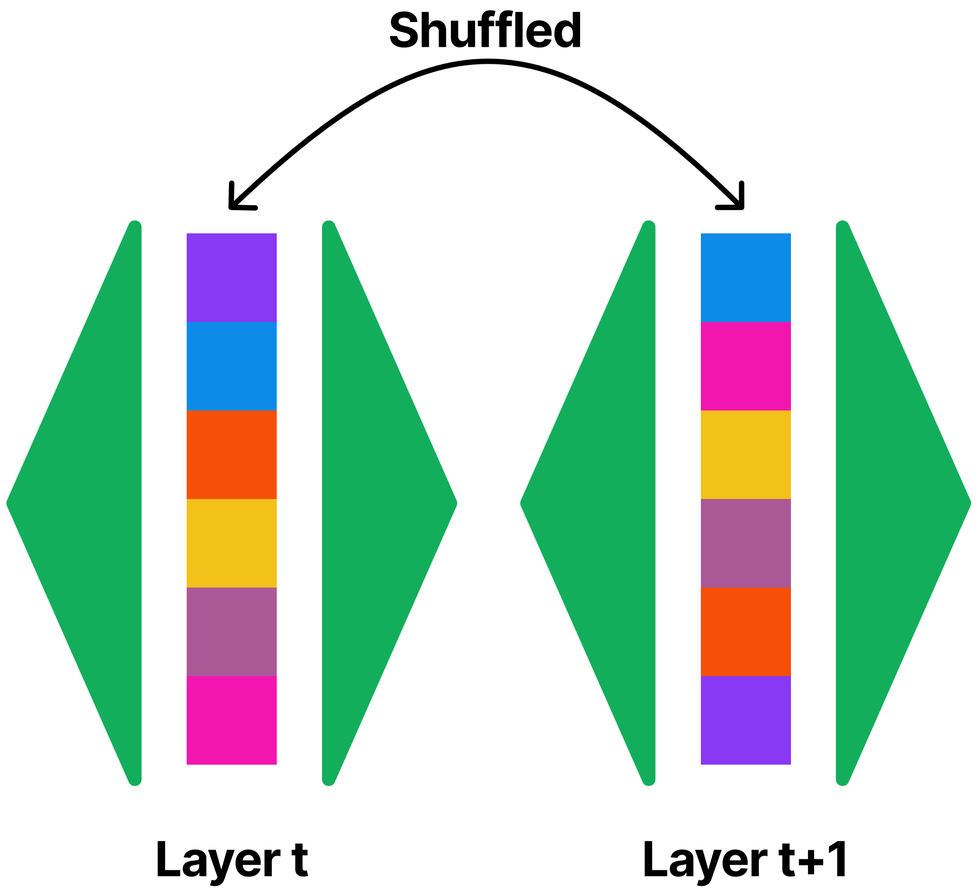
Hypothesis

- 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.

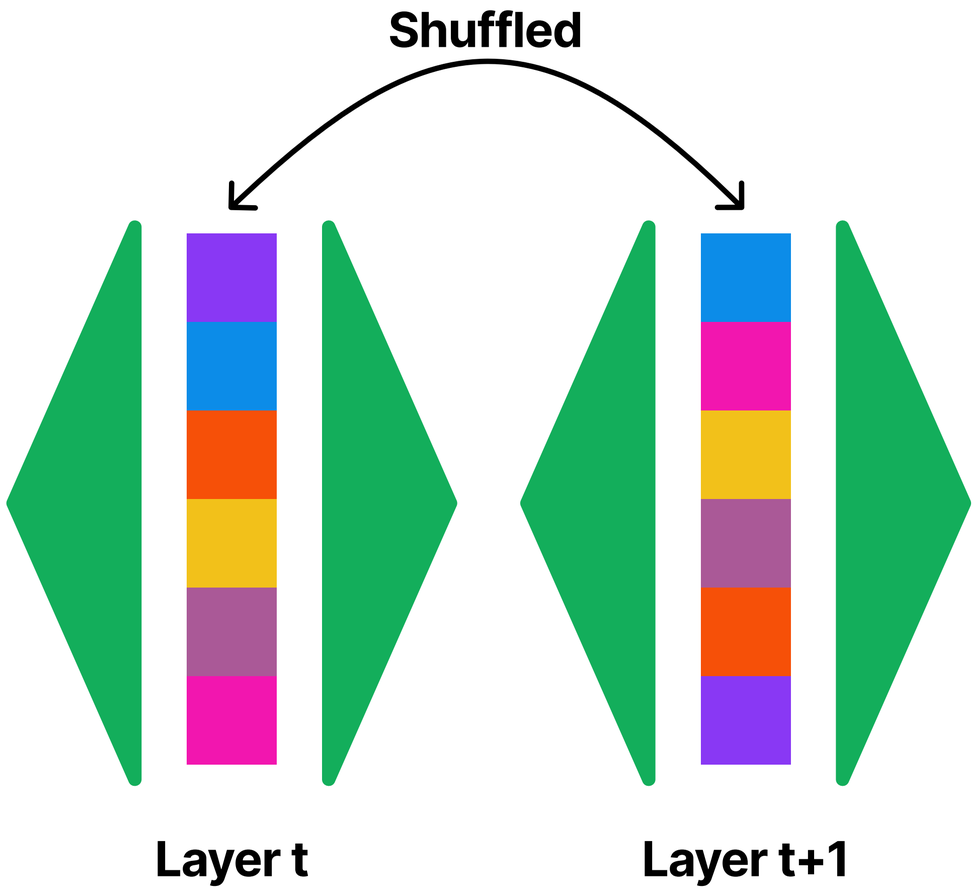


There is a problem with latents on different layers.



There is a problem with latents on different layers.

- Find correlation between features

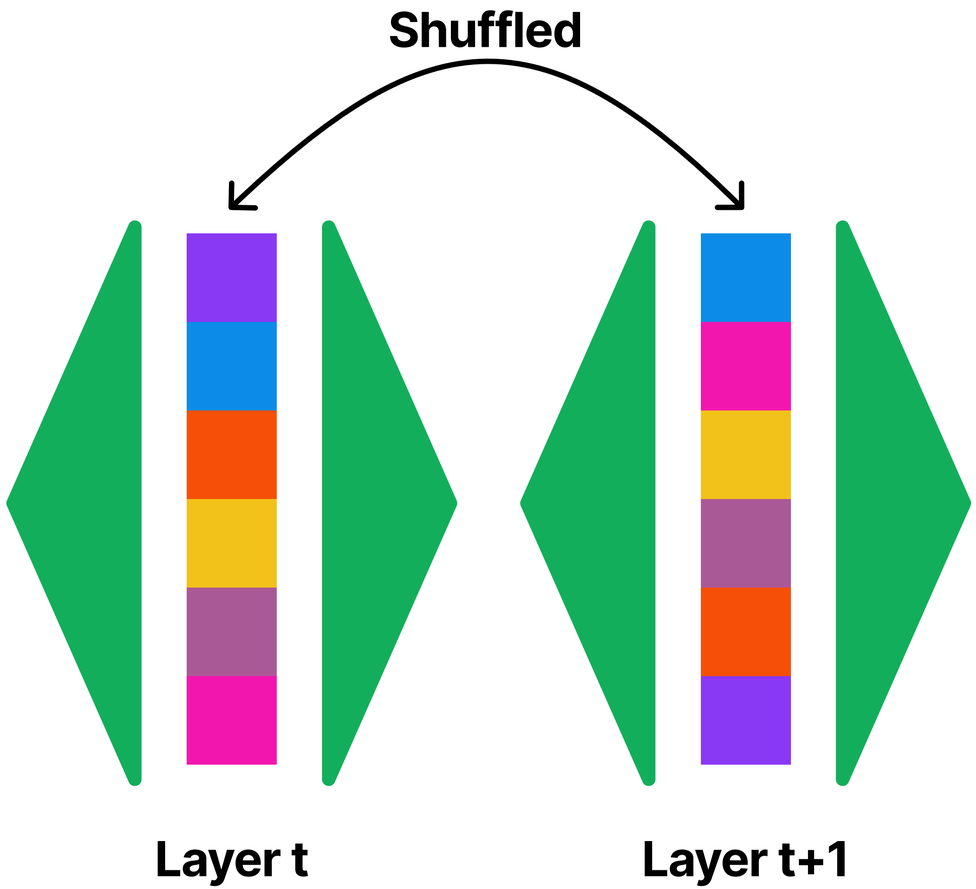


There is a problem with latents on different layers.

- Find correlation between features

(requires data and computational resources)

- Use weights to find permutation



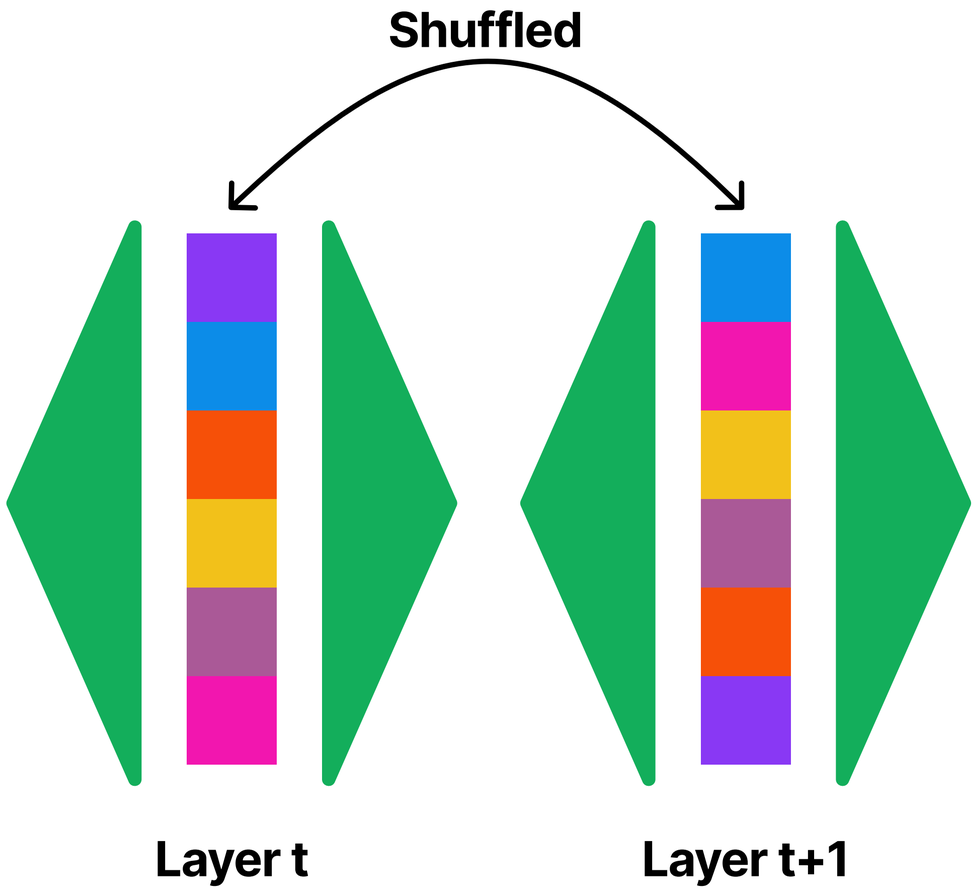
There is a problem with latents on different layers.

- Find correlation between features

(requires data and computational resources)

- Use weights to find permutation

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



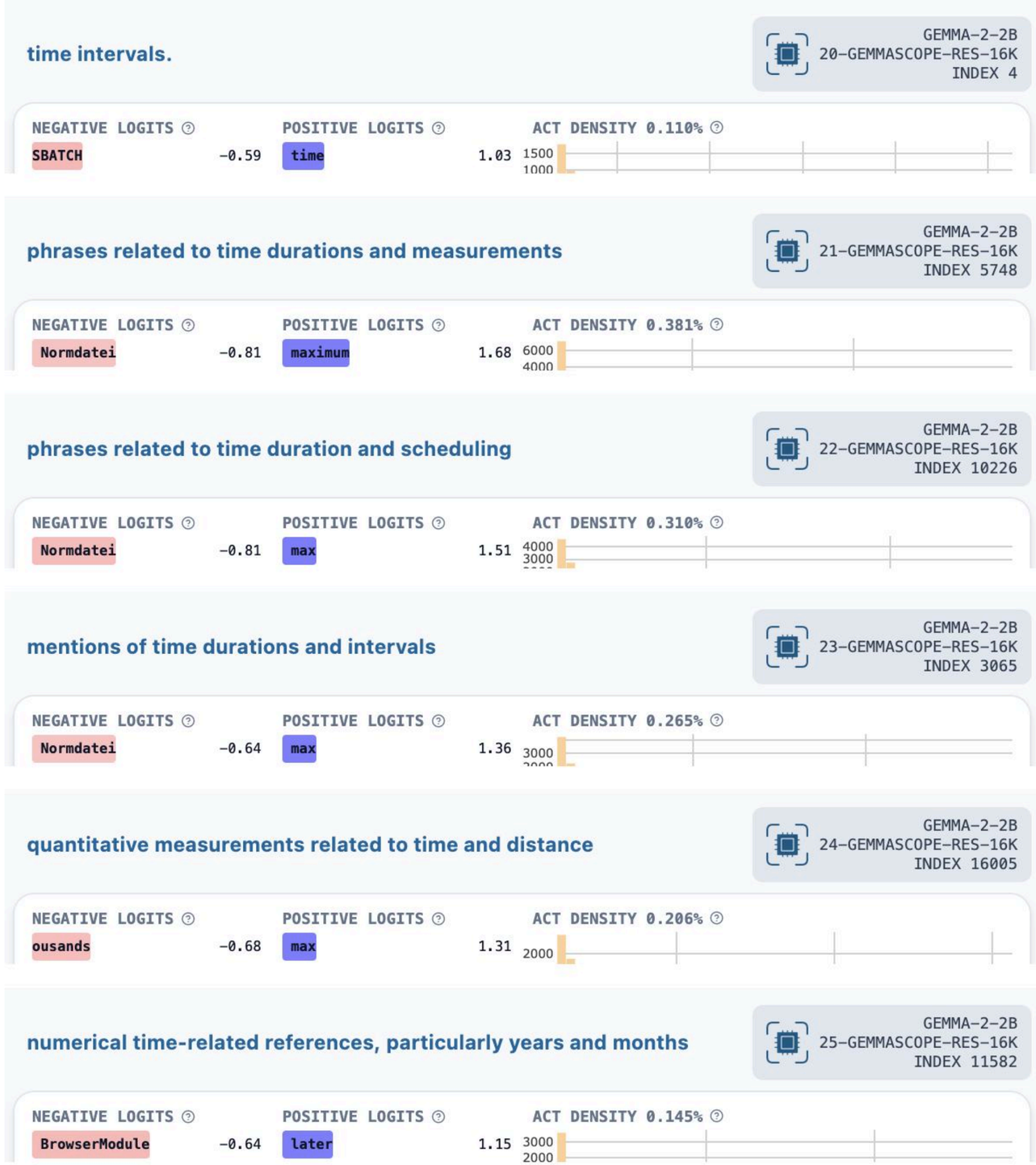
Obtained example

- Find correlation between features

(requires data and computational resources)

- Use weights to find permutation

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



Let’s validate

- Find correlation between features

(requires data and computational resources)

- Use weights to find permutation

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

Check out the full version here.