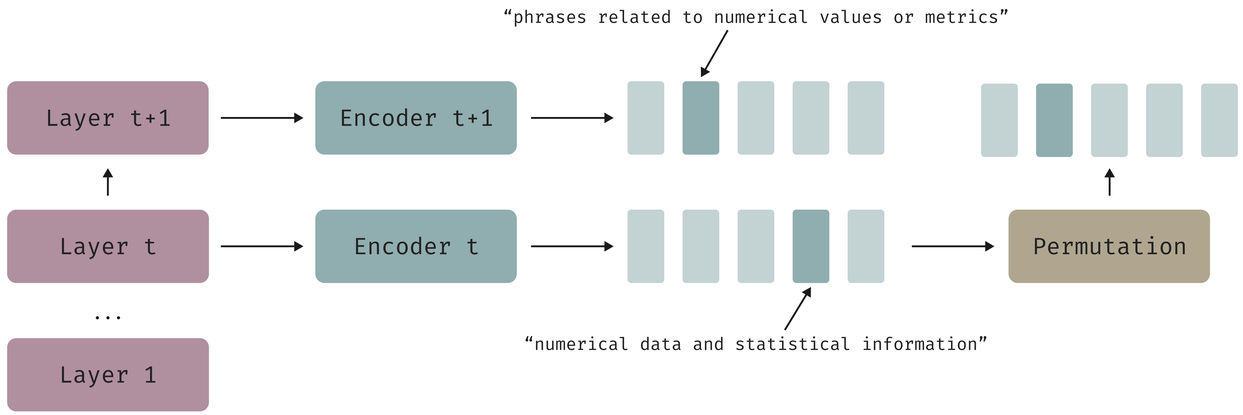
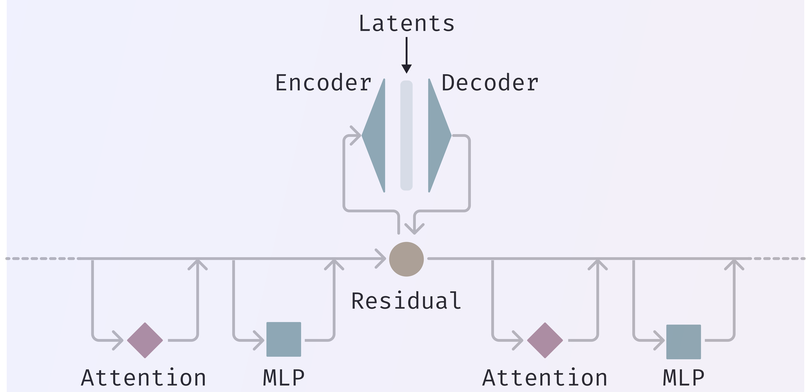
Mechanistic Permutability:
Match Features Across Layers.

Author Layer — **Affilation Layer**

Nikita Balagansky — T-Tech
Ian Maksimov — MIPT
Daniil Gavrilov — HSE University

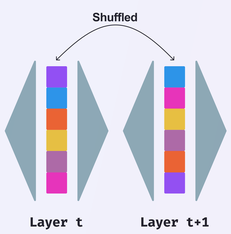


SAE



Sparse Autoencoders for LLMs use sparsity constraints to decompose model activations into interpretable components, isolating specific features or concepts. By penalizing dense activations, SAEs ensure only relevant neurons fire for a given input, enabling researchers to map these activations to human-understandable behaviors (e.g., factual knowledge or reasoning errors).

Multi-layer features



While SAEs extract interpretable features from individual layers, features across layers may be permuted or vary in scale due to differences in activation thresholds and hidden state dynamics. Existing methods do not provide data-free solution for these task. This work addresses the problem of aligning SAE features across layers by accounting for both permutation and scale variations, enabling accurate tracking of feature transformations through the network.

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.

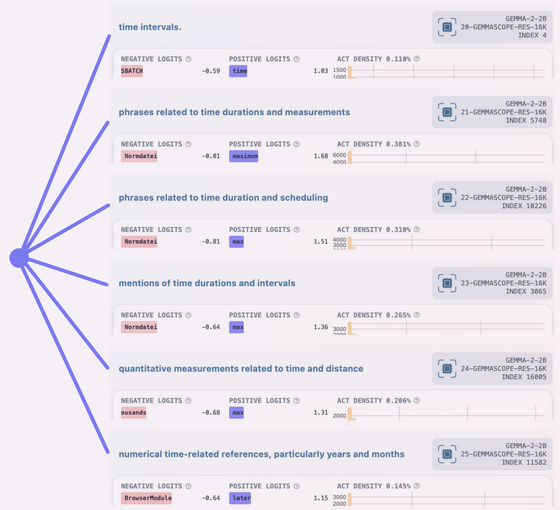
Method

Find permutation of the featres indexes P that satisfies:

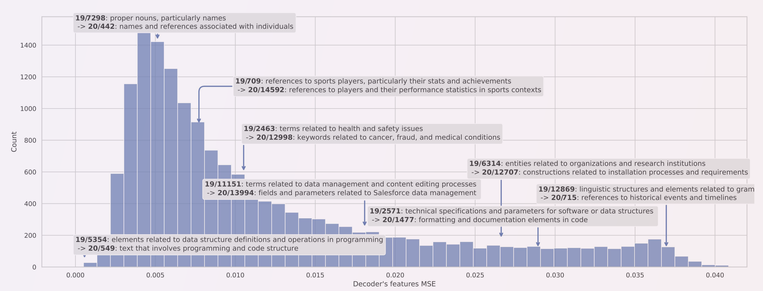
$$\arg\min_{\boldsymbol{P}\in\mathcal{P}_F} \sum_{i=1}^{d} \left\| \boldsymbol{W}^{(A)}_{\mathrm{dec}_{i,:}} - \boldsymbol{P}\boldsymbol{W}^{(B)}_{\mathrm{dec}_{i,:}} \right\|^2$$

Obtained Path

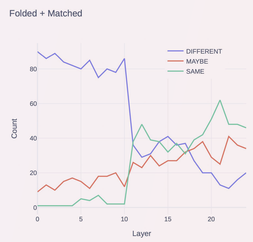
Same semantic on different layers!



MSE of the Matched Features' Embeddings

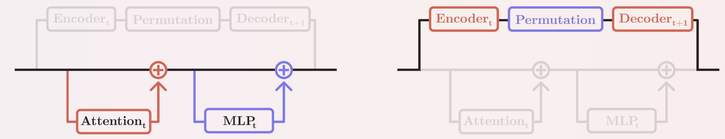


LLM Evaluation

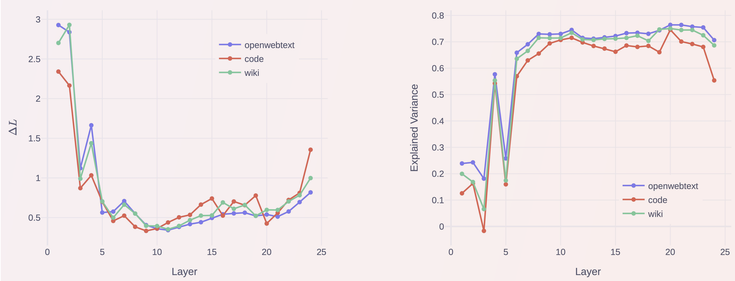


To validate our approach, we collected pairs of descriptions from matched features. We then used an external LLM (GPT-4o mini) to compare Neuronpedia descriptions of these feature pairs, categorizing them as "**SAME**" (semantically identical), "**MAYBE**" (plausibly related), or "**DIFFERENT**" (unrelated).

Permutation Allows Us To Create Transcoders



Performance of the Resulting Transcoders



Full Paper: