Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Mapping Feature Drift: Cross-Layer Correspondence and Polysemanticity in Qwen2.5-3b

Sahil Adit

sahilpradipadit@aissmscoe.com

AISSMS College Of Engineering https://orcid.org/0009-0003-3325-2693

Research Article

Keywords: Sparse Autoencoders , Mechanistic Interpretability , Cross-Layer Correspondence , Polysemanticity , Feature Disentanglement

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11706467/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

MAPPING FEATURE DRIFT: CROSS-LAYER CORRESPONDENCE AND POLYSEMANTICITY IN QWEN2.5-3B

Sahil Adit
AISSMS College Of Engineering
sahilpradipadit@aissmscoe.com

ABSTRACT

Sparse Autoencoders (SAEs) are one of the most significant method in Mechanistic Interpretability, offering a means to break down the internal activations of Large Language Models (LLMs) into features that are easier to understand [1, 2]. The key challenge is to study how these latents works inside these complex models.

We explore feature representations in the Qwen2.5-3B model [3] by examining two complementary aspects: cross-layer feature correspondence and polysemantic activation behavior. We use Hungarian match to analyse feature activation patterns across 36 transformers [4].

Our initial experiment discovers sparse and depth-dependent relations across layers, with just 17 of 8,960 potential matches. The second experiment explores the polysemantic behaviour of Qwen2.5 3B model. An important observation arises from Layer 33, where Feature F374 demonstrates significant activations across sentences related to various religions, philosophical traditions, including those from Christian, Jewish, Catholic, and Chinese philosophical backgrounds. Its significantly reduced activations in chosen technology and sports control instances offer additional proof of its contextual selectivity.

Collectively, these experiments offer insights into the evolution of cross-layer features and the contextual variety of internal representations in transformer networks, enhancing understanding of feature organization in large language models.

Our source code is publicly available at `https://github.com/sahiladit/MFD_Qwen2.5_3B`.

Keywords Sparse Autoencoders · Mechanistic Interpretability · Cross-Layer Correspondence · Polysemanticity · Feature Disentanglement

1 Introduction

The aim of mechanistic interpretability is to understand how neural networks function internally. We investigate layer-by-layer representation and processing. In the past, scientist have studied individual neurons to understand their behavior. A single neuron can react to several unrelated inputs and a single concept may also be shared by several neurons. Sparse Autoencoders (SAEs) have become a significant method for investigating interpretable features inside language models. These SAEs are trained on the internal activations of the model. This makes it possible to investigate language model's internal representations outisde of individual neurons.

Dictionary learning can be used to find additional interpretable features in language models, as early work by Bricken et al. showed. Cunningham et al. investigated recovery of highly interpretable features from neural network activations using Sparse Autoencoders [1, 2]. Different characteristics such as active features, quality of reconstruction affect how we understand the learn representation [5, 6].

Multi-layer SAE's are studied to understand how feature behave as information passes through the residual stream and how they active at various transformer depths [7]. In an effort to find connections between features learned at various layers, other research has looked into cross-layer feature matching and correspondence [4]. We try to study how these features behave inside the transformer layer but deciphering SAE representations becomes more difficult due to feature

splitting and absorption. A feature may represent several other features i.e. it may fire for multiple types of inputs, it is called Polysemanticiy.

We investigate these questions using Qwen2.5-3B Base, a transformer-based language model with 36 transformer blocks [3]. We train independent Top-K Sparse Autoencoders for the residual stream activations of each transformer layer. Each SAE learns a separate feature dictionary, allowing us to examine the activation behaviour of features at different depths of the model.

Our first experiment examines features across layers. We try to obtain feature activation across neighboring layers. We obtain feature activation profiles from a specialized linguistic dataset and analyze feature across successive transformer layers through correlation-based matching and the Hungarian algorithm. We Furthermore utilize permutation-based statistical testing with a global maximum-statistic correction to address multiple comparisons among feature pairs and layer transitions.

Based on our cross-layer analysis, we further investigate how SAE features behave across the model's depth, particularly whether individual features respond to multiple distinct concepts. Our second experiment focuses on polysemanticity by examining the natural-language contexts that strongly activate individual features and analysing their semantic diversity. Through these experiments, we aim to understand how features evolve across transformer layers and whether polysemantic representations progressively transform into more specialized and distinguishable features.

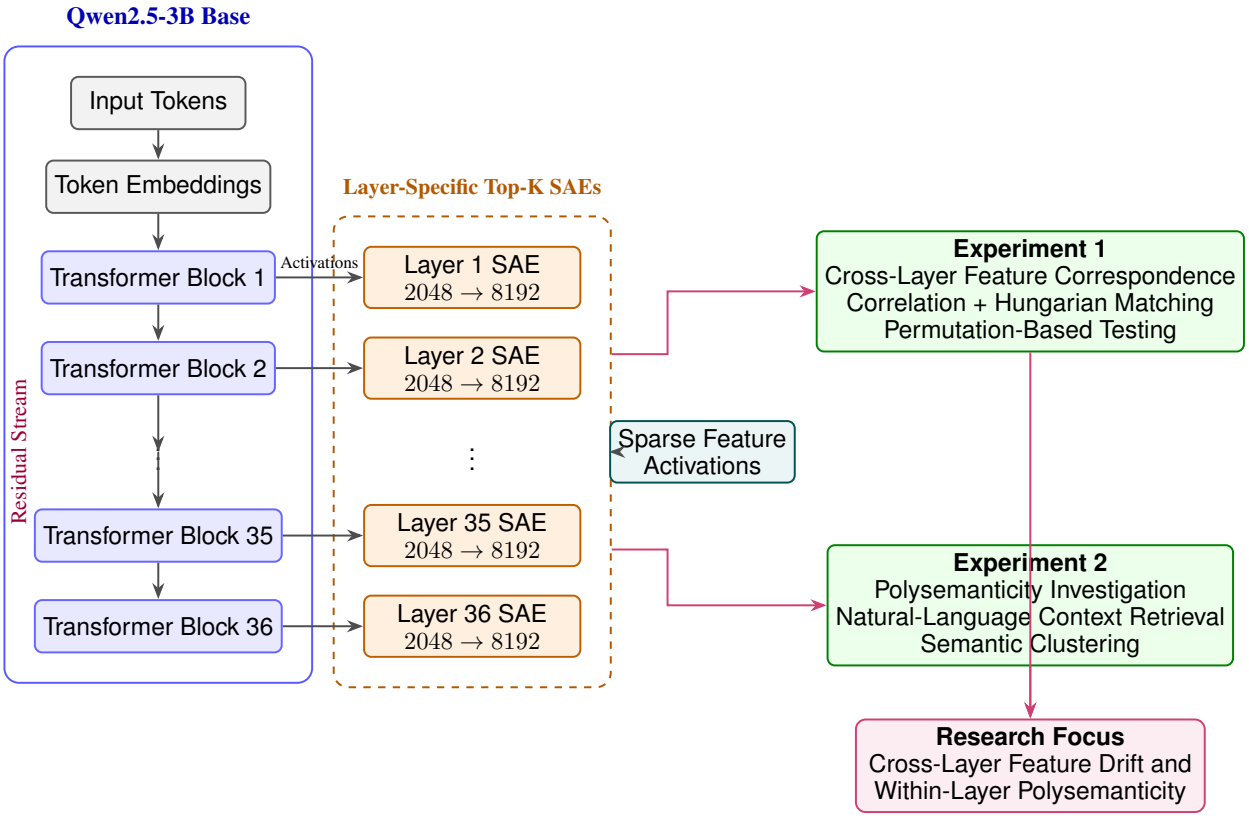


Figure 1: Overall research architecture showing the extraction of layer-wise residual stream activations using independent Top-K Sparse Autoencoders and the two experimental directions: cross-layer feature correspondence and polysemanticity analysis.

2 Background

2.1 Sparse Autoencoders

Sparse Autoencoders (SAEs) are one of the most important tools in Mechanistic Interpretability. They help us understand the internal behaviour of neural network layers by decomposing complex activation patterns into a set of simpler, more interpretable features. Instead of directly analysing individual neurons, SAEs learn a larger set of feature directions, commonly called latents, where only a small number of features are active for a given input. This approach is motivated

by dictionary learning and the idea that neural network activations may represent multiple underlying features through superposition [1, 2]. Latents are generally tensors or features.

An SAE mainly consists of an Encoder and a Decoder. The encoder takes the activation vector from the transformer layer and projects it into a higher-dimensional feature space. We have used ReLU as our activation function, it has been observed using Jump ReLU over simple ReLU gives better results. A sparsity constraint is then applied so that only a limited number of features remain active. The decoder uses these sparse feature activations to reconstruct the original input activation [2, 6].

The encoder and decoder can be mathematically represented as:

$$\mathbf{f}(x) = TopK\left(ReLU(W_{\text{enc}}x + b_{\text{enc}})\right) \tag{1}$$

$$\hat{x} = W_{\text{dec}}\mathbf{f}(x) + b_{\text{dec}} \tag{2}$$

Here, $x \in \mathbb{R}^d$ represents the original activation vector, while $\mathbf{f}(x) \in \mathbb{R}^F$ represents the sparse feature activations. W_{enc} and W_{dec} are the encoder and decoder weight matrices, respectively, and b_{enc} and b_{dec} are their corresponding biases. The reconstructed activation is represented by $\hat{x}$. In our experiments, we use Top-K Sparse Autoencoders with an input dimension of 2048, a feature dimension of 8192, and a Top-K value of 32. This means that each activation is represented using a maximum of 32 active features out of 8192 learned features.

Reconstructing the initial activation while preserving sparse feature representations is the primary goal of an SAE. This enables us to investigate which features are activated for various inputs and how their activation behavior varies throughout transformer levels. In order to examine feature correspondence and polysemanticity across the residual stream, we train an independent SAE for each of the 36 transformer blocks of Qwen2.5-3B Base.

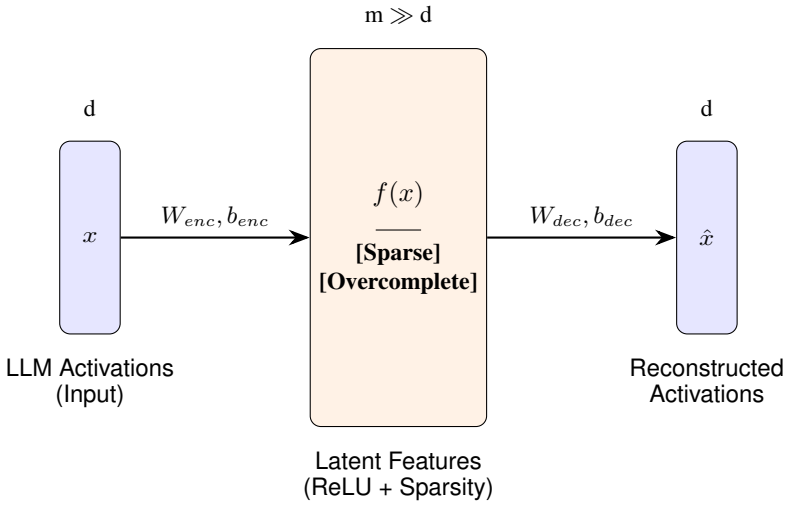


Figure 2: Architecture of a Top-K Sparse Autoencoder. The encoder maps the input activation into an overcomplete latent representation, applies ReLU and Top-K sparsity, and reconstructs the original activation through the decoder.

2.2 Residual Stream Activation Extraction

To investigate the internal representations across the depth of the transformer, we extract residual stream activations from each of the 36 transformer blocks of Qwen2.5-3B Base [3]. The model is loaded using a fixed Hugging Face revision to ensure reproducibility. For an input batch of token sequences, the output of transformer block l is represented as

$$H_l \in \mathbb{R}^{B \times T \times d_{\text{model}}}, \tag{3}$$

where B denotes the batch size, T represents the sequence length, and $d_{\text{model}} = 2048$ is the residual stream dimensionality.

Forward hooks are registered on transformer layers to record the output of each decoder block. The outputs of the transformer block RMSNorm operation is matched with the retrieved activations that results in the activation of residual stream.

The extraction process is performed independently for all 36 transformer blocks. Let H_l denote the activation tensor obtained from block l, where $l \in \{1, 2, \ldots, 36\}$. The activation vector corresponding to token position t at layer l is therefore given by

$$\mathbf{h}_{l,t} \in \mathbb{R}^{2048}. \tag{4}$$

For activation collection, we use the WikiText-103 raw training corpus [8]. WikiText-103 is tokenized using the tokenizer associated with Qwen2.5-3B and divided into fixed-length sequences of 256 tokens. The model processes these sequences, and the resulting activation tensors are stored separately for each transformer layer. This preserves the correspondence between token positions and their respective layer-wise representations.

The extracted activations are used as an input to the Sparse Autoencoder in the training process. We trained an independent Top-K Sparse Autoencoder to study different depths of the residual stream.

2.3 Dataset Preparation

The main text source for gathering residual stream activations and training the layer-specific Sparse Autoencoders is the WikiText-103 raw training corpus. A popular tool for assessing language modeling techniques is WikiText-103, a large-scale dataset of high-quality Wikipedia articles [8].

The tokenizer for the Qwen2.5-3B Base model [3] is used to tokenize the raw text. The recovered token sequences can directly reflect the original corpus text because special tokens are not added during tokenization. Fixed-length sequences of 256 tokens are created from the resultant token stream.

In the activation collection experiment, the assembled dataset includes 48,640 tokens, which are 190 sequences of 256 in length. Each sequence is handled separately by the language model to generate the associated residual stream activations within each transformer block. The predetermined sequence length ensures a uniform input structure for activation retrieval and later SAE training. We use the generated token sequences are utilized to form the layer-wise activation datasets, with every activation vector maintaining its link to the specific token position and transformer layer.

2.4 SAE Training and Reconstruction Evaluation

To examine the feature representations acquired at various depths of the transformer, we train a separate Sparse Autoencoder (SAE) for each of the 36 transformer layers. Every SAE obtains the relevant residual stream activations derived from its linked layer. This layer-by-layer training approach by layer enables the acquired features to embody the activation patterns unique to every transformer level.

Each SAE consists of an encoder and a decoder. Given an input activation vector $\mathbf{x} \in \mathbb{R}^{2048}$, the encoder maps the input into an expanded latent representation using

$$\mathbf{f}(\mathbf{x}) = TopK\left(ReLU\left(W_{\text{enc}}\mathbf{x} + \mathbf{b}_{\text{enc}}\right)\right), \tag{5}$$

where $W_{\text{enc}} \in \mathbb{R}^{8192\times 2048}$ denotes the encoder weight matrix, $\mathbf{b}_{\text{enc}}$ represents the encoder bias, and $ReLU$ is the rectified linear unit activation function. The $TopK$ operation retains only the 32 largest activations and sets all remaining activations to zero. Thus, each input activation is represented by at most 32 active features in an 8192-dimensional latent space.

The decoder reconstructs the original activation vector from this sparse latent representation:

$$\hat{\mathbf{x}} = W_{\text{dec}}\mathbf{f}(\mathbf{x}) + \mathbf{b}_{\text{dec}}, \tag{6}$$

where $W_{\text{dec}} \in \mathbb{R}^{2048\times 8192}$ is the decoder weight matrix and $\mathbf{b}_{\text{dec}}$ is the decoder bias. The reconstructed activation $\hat{\mathbf{x}}$ approximates the original input activation $\mathbf{x}$.

The training objective minimizes the reconstruction error between the original and reconstructed activation vectors. We express the reconstruction loss using the mean squared error (MSE):

$$\mathcal{L}_{\text{recon}} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2 , \tag{7}$$

where N denotes the number of activation vectors used for training. The Top-K activation constraint enforces sparsity in the latent representation, restricting the number of active features for each input.

We train 36 independent SAEs, one for each transformer block. Each autoencoder uses an input dimension of 2048, a latent dimension of 8192, and a Top-K sparsity parameter of $K = 32$. The last layer gives a common framework which examines the leaned feature representations across transformer depth.

Reconstruction Evaluation. Once the training is done, we checl the quality of reconstruction by comparing the original residual stream activations to their respective SAE reconstructions. The value of the reconstruction loss indicates how precisely the sparse latent representation maintains data present in the initial activation vectors.

We also measure the latent sparsity by analyzing the count of non-zero activations withing the encoded representations. In Top-K activations every input activation contains a maximum of 32 active latent features from a total of 8192 feasible characteristics.

Reconstruction quality is an important factor in feature-oriented evaluations. Rebuilding error should be considered while analyzing feature activations and cross layer feature alignment.

3 Experiment I: Cross-Layer Feature Correspondence

3.1 Experimental Objective

First experiment, *Cross-Layer Feature Correspondence*, explains if characteristics acquired by separately trained Sparse Autoencoders (SAEs) demonstrate steady activation patterns throughout successive transformer layers. Although every SAE is trained separately, characteristics at various layers and depths may encode associated data or show comparable activation patterns. To understand how feature repsentations develop first we need to understand these relationships between the layers.

The main goal is to measure the relationship between SAE characteristics. at neighboring transformer layers and assess if the detected associations generalize beyond the information utilized to recognize them. Instead of presuming that features that have matching indices in different layers convey the same information, we examine their activation patterns across a shared set of input examples.

We take into account all 36 transformer layers, meaning we have 35 transitions between transformer layers. From every layer, 256 features are active. For each set of adjoining layers, we assess the similarity of the activation patterns with their chosen characteristics. The Hungarian method is subsequently employed to determine pairwise feature associations derived from the activation similarity matrix.

Pairing is carried out with validation data and the resulting feature combinations are then assessed on a separate, reserved test dataset to evaluate whether the established correspondences can be generalized. This separation guaranties that the assessment does not entirely depend on the data utilized to determine the relationships.

The experiment therefore addresses three central questions:

1. **Feature Correspondence:** To what extent do features in adjacent transformer layers exhibit similar activation patterns?
2. **Generalization:** Do feature correspondences identified on validation data remain consistent when evaluated on unseen test examples?
3. **Statistical Significance:** Are the observed cross-layer correspondences stronger than those expected under an appropriate permutation-based null model, after accounting for multiple comparisons?

Overall, this experiment establishes a quantitative baseline for studying feature relationships across transformer depth. It also provides the methodological foundation for subsequent experiments investigating feature drift, polysemanticity, and changes in learned representations across the network.

3.2 Feature Activation Profiles

To measure the relationship between characteristics across successive transformer layers, we depict each SAE feature through its activation profile across a shared series of regulated benchmark instances. A trigger profile reflects

the intensity of a specific characteristic's reaction to various inputs, establishing a foundation for assessing feature performance across depth of transformer.

We managed to evaluate 3072 instances, split them into 1536 validation samples and 1536 reserved test samples for benchmarking the model. The benchmark is categorized into three linguistic elements : the valence of adjectives, the existence of denial and grammatical voice. SAE activations are obtained at the adjective token range of every instance, permitting the examination to concentrate on the feature replies linked to the regulated linguistic differences.

Let $a_{\ell,j}^{(i)}$ denote the activation of feature j in transformer layer ℓ for benchmark example i. For a benchmark containing N examples, the activation profile of feature j at layer ℓ is defined as

$$\mathbf{p}_{\ell,j} = \left[a_{\ell,j}^{(1)}, a_{\ell,j}^{(2)}, \ldots, a_{\ell,j}^{(N)}\right], \tag{8}$$

where $\mathbf{p}_{\ell,j} \in \mathbb{R}^N$ represents the activation pattern of the feature across the evaluated examples.

In every transformer layer, the 256 most engaged qualifying features are chosen utilizing the validation dataset. The identical chosen feature IDs are then preserved for the reserved test assessment. This guarantees that characteristic selection is not independently replicated on the test split.

To measure the similarity between features in consecutive layers, we compute the Pearson correlation between their activation profiles. For a feature j in layer ℓ and a feature k in layer $\ell + 1$, the correlation is given by

$$r_{\ell,j,k} = \frac{\sum_{i=1}^{N}\left(a_{\ell,j}^{(i)} - \bar{a}_{\ell,j}\right)\left(a_{\ell+1,k}^{(i)} - \bar{a}_{\ell+1,k}\right)}{\sqrt{\sum_{i=1}^{N}\left(a_{\ell,j}^{(i)} - \bar{a}_{\ell,j}\right)^2}\sqrt{\sum_{i=1}^{N}\left(a_{\ell+1,k}^{(i)} - \bar{a}_{\ell+1,k}\right)^2}}, \tag{9}$$

where $\bar{a}_{\ell,j}$ and $\bar{a}_{\ell+1,k}$ denote the mean activations of the respective features over the evaluated examples.

As a result of correlations between pairs, a similarity matrix for each neighboring-layer shift is created. These matrices are the foundations of the Hungarian matching method, which establishes one-to-one feature correspondences.

This depiction allows the experiment to investigate whether the characteristics that display comparable activation patterns during validation persist in react comparably to unknown standard instances. Mainly, stimulation profile similarity metrics indicate statistical correlation rather than semantic similarity or causal correspondence between attributes.

3.3 Hungarian Matching

To determine feature matches between successive transformers. In the layers, we utilize the Hungarian algorithm, which addresses the linear assignment issue by determining an ideal one-to-one pairing between a pair of feature sets.

For every transition between adjacent layers, the Pearson correlation coefficients among the activation patterns of source-layer and target-layer characteristics are organized into a similarity matrix. Let ℓ and $\ell+1$ denote two consecutive transformer layers, each containing 256 selected candidate features. The resulting similarity matrix is defined as

$$\mathbf{C}^{(\ell)} \in \mathbb{R}^{256\times 256}, \tag{10}$$

where each entry is given by

$$C_{j,k}^{(\ell)} = corr\left(\mathbf{p}_{\ell,j}, \mathbf{p}_{\ell+1,k}\right). \tag{11}$$

The objective is to determine an assignment that maximizes the total correlation across all matched feature pairs. Formally, the assignment is expressed as

$$\max_{\mathbf{X}^{(\ell)}} \sum_{j=1}^{256}\sum_{k=1}^{256} X_{j,k}^{(\ell)} C_{j,k}^{(\ell)}, \tag{12}$$

subject to the one-to-one assignment constraints

$$\sum_{k=1}^{256} X_{j,k}^{(\ell)} = 1, \qquad \sum_{j=1}^{256} X_{j,k}^{(\ell)} = 1, \tag{13}$$

where $X_{j,k}^{(\ell)} \in \{0, 1\}$ indicates whether feature j in layer ℓ is assigned to feature k in layer $\ell + 1$.

Our in depth analysis revealed additional variation across the network. Early-layer transitions (1–10) exhibited a mean test correlation of 0.0537, while middle-layer transitions (11–24) achieved 0.0799. Late-layer transitions (25–35) showed a comparatively higher mean test correlation of 0.1332. However, this increase does not establish consistent feature preservation in deeper layers, as the overall test correlations remained substantially lower than their validation counterparts.

3.4 Statistical Validation

We conducted a permutation-based significance analysis to ascertain if the observed cross-layer correlations indicated statistically significant relationships rather than random associations. The Hungarian assignments obtained from validation were maintained as constant, while the structure of benchmark contrasts was retained throughout the permutation process. This enabled us to assess the statistical robustness of the aligned feature pairs without adding extra matching choices while testing. We performed 499 matched-block permutations and assessed the resulting correlation statistics for all 8,960 matched pairs and 35 adjacent-layer transitions. We utilized a global maximum-statistic adjustment, to address multiple comparisons. We noted the highest correlation among all assessed feature pairs, creating a null distribution that allowed for the evaluation of the strongest observed relationship. The 95th percentile of the generated global null distribution was 0.9724. No arrangement yielded a maximum correlation higher than the observed peak. Considering the 499 permutations, the lowest achievable empirical p-value was

$$p_{\min} = \frac{1}{499 + 1} = 0.002. \tag{14}$$

The result shows that the most significant correlation reached a global empirical p-value of 0.002 at the existing permutation resolution. This outcome should not be viewed as proof that every matched pair was statistically significant. If we talk about global correlation, just 17 out of the 8,960 candidate correspondences were statistically significant, which accounts for roughly 0.19% of all assessed pairs. The distribution of these important correspondences by depth was likewise uneven. Early-layer transitions (1–10) included 11 significant pairs out of 2,560 candidates, while no significant correspondences were found in the middle-layer transitions (11–24), which had 3,584 pairs. The late-layer transitions (25–35) included the other 6 notable pairs from 2,816 candidates. Therefore, the statistically validated associations were focused on the initial and final segments of the network, with none observed in the central area using the implemented correction method.

These results indicate that there is no extensive preservation of feature activation patterns across successive transformer layers. Even though the matching process found potential correspondences across the network, merely a tiny percentage stayed meaningful after applying the global statistical correction. Therefore, the findings endorse a sparse and depth-dependent pattern of inter-layer correspondence instead of a universal continuity of feature representations. Essentially, statistical significance demonstrates a notably robust alignment of activation profiles based on the assessed benchmark and permutation framework; however, it does not independently confirm semantic equivalence, inherence of causal traits, or a mechanism for splitting and merging.

3.5 Results and Analysis

Talking about the Results, we display the experimental results derived from assessing adjacent-layer feature correspondences in the residual stream of Qwen2.5-3B Base. The study shows the generalization of aligned feature activation patterns, their statistical importance, and the differences in correspondence strength at various network depths.

3.5.1 Overall Cross-Layer Correspondence

We assessed 8960 potential feature correspondences from these 36 transformer blocks which covered all of the 35 transitions between adjacent layers. We found that the correlation between validation data and the text data decreased significantly over the course of network layers. The average correlation among paired features dropped from 0.4663 in the validation set to 0.0892 in the held-out test set. Furthermore, the median test correlation was 0.0000, suggesting that at least 50This inconsistency determines that the matches found during validation were typically not consistent across

Table 1: Overall cross-layer feature correspondence results.

Metric	Result
Transformer blocks	36
Adjacent-layer transitions	35
Candidate feature correspondences	8,960
Mean validation correlation	0.4663
Mean held-out test correlation	0.0892
Median held-out test correlation	0.0000
Validation–test score correlation	0.2393
Globally significant correspondences	17 (0.19%)

Table 2: Cross-layer correspondence statistics across network depth.

Region	Transitions	Pairs	Val. r	Test r	Significant
Early	1–10	2,560	0.4930	0.0537	11
Middle	11–24	3,584	0.4520	0.0799	0
Late	25–35	2,816	0.4603	0.1332	6

the reserved evaluation data. Although we found some connected activation patterns in the matching process during validation, but their significantly lower test correlations suggest restricted generalization. The Pearson correlation for the validation and test pair scores was 0.2393. Nevertheless, because the observations contain examples, layers, and assignment constraints, this value ought to be understood in a descriptive manner rather than as proof derived from 8,960 independent observations.

3.5.2 Depth-Wise Analysis

To check the strength of cross-layer correspondence, we categorized the 35 transitions into 3 regions : early (transitions 1–10), middle (11–24), and late (25–35). The outcomes showed variations in both the mean test correlation and the spread of statistically significant associations. Early-layer transitions showed an average validation correlation of 0.4930, which fell to 0.0537 on the reserved test set. In the central area, the related values were 0.4520 and 0.0799. Transitions in the late layer (25-35) exhibited a mean validation correlation of 0.4603 and a relatively greater test correlation of 0.1332. All three areas showed a significant drop in validation performance. Overall, feature activations simply do not hold steady at any particular depth.

The central area showed no statistically significant relationships, even though it showed a little bit positive validation correlations. The early area represented 11 out of the 17 notable pairs, whereas the late area comprised the other six. This distribution indicates that statistically validated activation-profile correspondences are focused in the initial and final segments of the network under the tested benchmark. Nevertheless, the limited quantity of remaining pairs hinders us from understanding this pattern as a universal architectural characteristic.

3.5.3 Interpretation and Implications

The broader perspectives shows a significant disparity between the visible alignment of SAE characteristics in validation data and their performance on unseen instances. Although Hungarian algorithm recognized one-to-one pairings in each adjacent-layer transition, many of these pairings showed low test correlations. Just 17 out of the 8,960 candidate correspondences remained after global statistical correction. Their distribution in early and late transitions also shows that the surviving connections are sparse and depend on depth rather than being evenly spread across the network. These observations align with the idea that activation patterns of features vary as information moves through the residual stream. Nonetheless, the current experiment assesses the statistical relationship between activation patterns and does not directly determine the evolution of individual semantic features. Specifically, it fails to show that a feature gradually separates, divides into several features, or combines with another feature. The results thus offer an empirical foundation for exploring feature drift throughout transformer depth, while also encouraging a distinct analysis of individual features in natural-language situations. This differentiation is significant as cross-layer activation correspondence and within-layer polysemanticity illustrate different characteristics of learned representations and necessitate distinct experimental validation.

3.5.4 Limitations of the Results

The results must be understood taking into consideration controlled evaluation standard, which altered adjective valence and the presence of negation while assessing SAE activations at spans of adjective tokens. As a result, the findings may not universally apply to every linguistic context or semantic connection. Furthermore, the correlation of

activation profiles does not determine semantic equivalence or causal relationships. The reliability of feature activation measurements may also be affected by the quality of SAE reconstruction that depends on the layer. Ultimately, employing 499 permutations restricts the clarity of the empirical significance estimates to 0.002. Therefore, the main finding is confined to the assessed context: neighboring-layer SAE features show sparse, depth-related activation-profile alignment, with minimal indication of extensive retention of feature identity throughout successive transformer layers.

4 Experiment II: Investigating Polysemanticity

4.1 Experimental Objective

This experiment examines the existence of polysemanticity in the characteristics acquired by layer-specific sparse autoencoders (SAEs) within Qwen2.5-3B Base. The previous experiment investigated the consistency of feature activation patterns and residual stream activation through successive transformer layers, whereas this experiment centers on the semantic traits of individual features and analyzes if a single SAE feature can respond to various semantically different contexts. Polysemanticity denotes the occurrence where a neural unit reacts to various, possibly unrelated concepts or input patterns. A suggested explanation is *superposition*, where neural networks encode more features than the number of dimensions by incorporating several feature directions in the same representational space [9]. This polysemanticity complicates mechanistic interpretability: individual units rarely map cleanly to a single concept. To address this, sparse autoencoders project activations into a higher-dimensional space, isolating cleaner, more interpretable features from otherwise entangled representations [2].

Nonetheless, acquiring sparse feature activations does not automatically ensure that each learned feature corresponds to a single semantic concept. A feature can react to various subjects, language patterns, entities, or conceptual frameworks. Therefore, analyzing activation sparsity by itself is inadequate to determine if a learned feature is monosemantic or polysemantic. The main goal of this experiment is to explore if individual SAE features display various recognizable semantic connections when triggered by natural-language inputs. We analyze high-activation contexts linked to chosen features, recognize consistent semantic patterns, and assess whether noticeable differences indicate truly separate concepts or variations of a more extensive underlying theme. The research questions that direct the experiment are as follows:

1. Do individual SAE features exhibit activation across multiple semantically distinguishable natural-language contexts?
2. Can recurring activation contexts be organized into coherent semantic groups, and do these groups represent distinct concepts or related variations of a common theme?
3. Are the identified semantic associations consistent across different text corpora?
4. Do the observed feature activations have measurable functional effects when the corresponding SAE features are selectively intervened upon?

These are the questions that we are going to address in this experiment. We aim to distinguish Semantic diversity from superficial contextual variation including shared vocabulary, named entities, formatting patterns, and broad topical similarity. The investigation hence treats polysemanticity as a hypothesis requiring contextual and functional evidence.

4.2 Identification of Candidate Polysemantic Features

The early phase of Experiment 2 concentrated on recognizing SAE features that demonstrated possibly varied activation behavior in different textual contexts. We did not assume the feature was polysemous from its activation strength, rather we explored the contextual variation of its activation patterns. Instead of presuming that a feature was polysemous only based on its activation strength, we explored the contextual variation of its activation patterns to pinpoint candidates for more thorough examination. We examined the 36 layer-specific Sparse Autoencoders (SAEs) trained on the residual stream of Qwen2.5-3B. Every SAE contained 8,192 latent characteristics, utilizing a Top-K activation approach ($K = 32$). We chose 16 features from each layer, which lead to a cumulative total of 576 potential features throughout the model.

Context Retrieval. For every chosen feature, we gathered textual contexts connected to its most significant activations from the WikiText-103 corpus. The process of extracting activations utilized the residual stream representations linked to the specific transformer layers. For each feature, the highest activation context from every 256-token sequence was preserved for further analysis. This process generated a preliminary set of 23,040 contextual examples. After removing the duplicates, 23,036 distinct contexts were preserved. This information formed the foundation for exploring the linguistic and semantic variety.

Candidate Classification. The chosen features were then analyzed based on their activation and contextual evidence. Based on the analysis the features were characterized into three distinct groups :

- **Strong candidates:** Features exhibiting comparatively strong evidence of activation across diverse contextual patterns and warranting further investigation.
- **Provisional candidates:** Features displaying some degree of contextual diversity, but for which the available evidence was insufficient for stronger classification.
- **Unsupported candidates:** Features for which the retrieved contexts did not provide sufficient evidence of meaningful semantic diversity.

The preliminary classification revealed 10 robust candidates, 69 tentative candidates, and 497 unsubstantiated candidates from the 576 examined features. These categories were utilized to prioritize further contextual and semantic analyses instead of determining conclusive polysemanticity. Furthermore, 12,266 features were identified as displaying widespread activation patterns throughout the extensive feature analysis. Such patterns were approached with caution since widespread activation by itself does not confirm the existence of various semantic functions. The candidate set obtained served as a foundation for the following contextual activation assessment and semantic grouping experiments. Specifically, the most promising candidates were kept for further examination to ascertain if their activation contexts reflected distinct semantic domains or variations of a more general underlying concept.

4.3 Contextual Activation Analysis

To explore the contextual behavior of the chosen SAE features, we tested the textual contexts linked to their most significant activations. The aim was to investigate if separate characteristics reacted to various semantic and linguistic environments instead of being limited to one subject. The obtained contexts were analyzed for every feature in relation to their respective activation values. This enabled us to recognize repeated patterns and contrast the categories of content linked to specific features. The analysis showed that certain features displayed activation across various contextual categories, such as historical, cultural, religious, and literary content. Nonetheless, activation among various topics does not automatically suggest separate semantic roles, as these contexts might have a common conceptual connection.

Therefore, the contextual observations were utilized to recognize possibly varied activation patterns and direct the following semantic clustering analysis. This offered a qualitative foundation for exploring if the noted contextual variation indicated identifiable semantic categories.

4.4 Semantic Clustering and Stability Analysis

To analyze the semantic variation across retrieved activation contexts, we clustered contexts with similar textual patterns. We wanted to test whether the contexts linked to a given SAE feature formed distinct sub-groups or simply reflected variations of a broader semantic theme.

We first vectorized the text using TF-IDF, reduced its dimensionality with Truncated SVD, and then ran K-Means across multiple cluster counts and random seeds. To evaluate cluster quality and stability, we computed Silhouette Scores and tracked consistency across initialization seeds. This allowed us to measure both how cleanly the contexts separated and whether the groupings held up across runs.

We treated these clustering results as exploratory rather than definitive proof of separate semantic meanings. In particular, we evaluated high clustering scores alongside cluster sizes and raw text examples to avoid mistaking tiny, outlier clusters for genuine semantic divisions.

4.5 Targeted Investigation of Layer 33: Feature F374

While exploring features in Layer 33, feature F374 caught our eye. It offered a great test case: could one SAE feature reliably fire across very different religious and cultural settings?To investigate, we used the base Qwen2.5-3B model and its Layer 33 SAE (2,048 input dimensions, 8,192 latents, and Top-K activation with $K = 32$). We pulled the residual stream states from Layer 33, fed them into the encoder, and applied ReLU and Top-K filtering to get the final activation values.

Contextual Investigation. To evaluate F374, we tested six specific texts. Four focused on distinct religious, historical, and philosophical traditions: Christian eschatology, classical Chinese philosophy and literature, Jewish religious organizations, and Christian church practices. The remaining two came from sports and technology to serve as negative controls. By testing across these areas, we could see whether F374 was responding to a broad, cross-religious theme rather than overfitting to one particular doctrine.

Activation Results. Table 3 presents the activation measurements obtained for the six selected contexts. Raw maximum activation represents the highest encoder activation observed within a sample before Top-K selection. The number of active tokens indicates how many tokens produced a positive activation for F374. Top-K active tokens represent the number of tokens for which the feature remained active after the SAE's Top-K selection.

Table 3: Activation behaviour of Layer 33, Feature F374, across different contexts.

Context	Raw Max.	Raw Active	Top-K Active
Christian eschatology	60.03	20/50	3/50
Chinese philosophy and literature	63.59	26/30	13/30
Jewish religious institutions	61.37	19/20	9/20
Christian institutional activities	46.74	15/20	5/20
Technology (control)	13.23	5/18	1/18
Sports (control)	8.78	7/14	0/14

Across the test inputs, F374 fired far more strongly on the religious and philosophical texts than on the controls. It peaked at 63.59 on Chinese philosophy and literature, followed closely by 61.37 on Jewish religious institutions, with similarly high values for both Christian eschatology and institutional practice. The sports and tech controls stayed near baseline. Because F374 responded consistently across Christianity, Judaism, and Chinese thought, it does not seem tied to any single faith's vocabulary or doctrine.

Interpretation. These patterns point to polysemantic behavior in Layer 33: a single SAE feature fires strongly across distinctly different religious and cultural settings. Because activations dropped off sharply in the tech and sports controls, this response is unlikely to be background noise. Rather than tracking a single religious doctrine, F374 appears to capture a broader abstraction around religious and philosophical discourse. However, our tests do not prove that each context triggers a distinct semantic sub-function, and mapping the exact internal structure of this feature will require further work. Overall, F374 provides a clear empirical example of polysemantic feature dynamics inside Qwen2.5-3B's residual stream.

4.6 Results and Discussion

In our second experiment, we evaluated the contextual diversity of SAE features through a four-stage pipeline: candidate identification, contextual activation analysis, clustering, and targeted feature inspection. Out of 576 initial features, our screening labeled 10 as strong candidates, 69 as potential candidates, and 497 as unsupported.

We then benchmarked the 79 combined candidates against 79 matched control features using 500 WikiText articles and 500 C4 documents. Candidate features achieved an average silhouette score of 0.05831 compared to 0.05852 for controls, with corresponding stability scores of 0.87656 and 0.87087. Ultimately, the candidates offered no measurable improvement in clustering quality over matched controls, and none replicated reliably across corpora under our validation criteria.

Our analysis of Layer 33, Feature F374, showed strong activations across texts spanning Christian eschatology, Jewish organizations, Catholic communities, and Chinese philosophy, while the tech and sports controls barely registered. Peak activation occurred on the Chinese philosophy and literature passage, confirming that F374 fires across several distinct contextual settings.That said, because this targeted inspection relied on only six curated prompts, it cannot tell us how widespread this pattern is or prove that the feature serves separate semantic functions. Similarly, our ablation results remain exploratory. Ablating candidate features yielded an average Jensen-Shannon divergence of 3.19×10^{-5} compared to 5.19×10^{-6} for controls, and only one out of 30 candidate ablations altered the top-1 token prediction (versus zero of 20 for controls). Without running these interventions on strictly matched inputs, we cannot isolate candidate-specific causal effects.

In short, our data hints at contextual diversity within SAE features—particularly Feature F374 in Layer 33—but the clustering and cross-corpus checks tell a more cautious story. We found no consistent evidence of candidate-specific semantic separation. F374 clearly activates across different settings, but whether this points to distinct, autonomous sub-functions or just an overarching, generalized concept remains an open question.

5 Conclusion

We investigated sparse autoencoder (SAE) feature dynamics in the residual stream of Qwen2.5-3B, focusing on cross-layer feature alignment and polysemantic activations across two complementary experiments.

In our first experiment, we tracked feature correspondence across all 36 transformer layers. While validation splits showed seemingly strong correlations, these relationships largely evaporated on held-out test data—mean correlations

fell from 0.4663 down to 0.0892. After applying a global multiple-testing correction, only 17 of the 8,960 candidate pairings remained statistically significant, demonstrating that persistent cross-layer feature alignment is remarkably rare.

Our second experiment tested for polysemantic activations across 576 candidate SAE features. The clearest example surfaced at Layer 33 with Feature F374. It activated heavily on passages from Christian eschatology, Jewish organizations, Catholic community life, and Chinese philosophy (across Buddhist, Daoist, and Confucian texts), while remaining largely quiet on sports and tech controls. This cross-domain firing demonstrates that SAE features do not always isolate single concepts—instead, a single feature can respond across several distinct semantic boundaries.

This work provides an empirical look at cross-layer feature drift and polysemanticity in transformer representations. Our findings highlight why quantitative activation tracking must be paired with contextual inspection and rigorous controls—isolated metrics alone miss how large language models actually organize internal concepts.

References

[1] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

[2] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

[3] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[4] Nikita Balagansky, Ian Maksimov, and Daniil Gavrilov. Mechanistic permutability: Match features across layers. In *International Conference on Learning Representations*, 2025.

[5] Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, 2025.

[6] Alireza Makhzani and Brendan Frey. k-sparse autoencoders. In *International Conference on Learning Representations*, 2014.

[7] Tim S. Lawson, Lucy Farnik, Conor Houghton, and Laurence Aitchison. Residual stream analysis with multi-layer saes. In *International Conference on Learning Representations*, 2025.

[8] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

[9] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.