ALIGNING SENTENCE EMBEDDINGS TO HUMAN CONCEPTS VIA SPARSE AUTOENCODERS

Wonseok Shin, Songkuk Kim
Yonsei University
{wonseok.shin,songkuk}@yonsei.ac.kr

ABSTRACT

Dense sentence embeddings are fundamental to modern Retrieval-Augmented Generation (RAG) systems but suffer from a lack of interpretability due to feature superposition. This opacity hinders the alignment of retrieval processes with human intent, as the entangled representations are difficult to analyze or control. In this work, we propose a method to disentangle the dense representations of sentence transformers (e.g., E5) into human-interpretable concepts using Top-k Sparse Autoencoders (SAEs). We demonstrate that these disentangled features align with specific semantic, syntactic, and pragmatic categories. Furthermore, we introduce an activation steering mechanism that allows for precise intervention in the retrieval process. By clamping specific latent features, we show that it is possible to re-rank search results to better align with user constraints without retraining the backbone model. Our findings suggest that SAE-based decomposition offers a viable path toward transparent and steerable neural information retrieval.

1 INTRODUCTION

Neural sentence embeddings serve as the backbone of modern retrieval-centric applications, including Retrieval-Augmented Generation (RAG) pipelines (Lewis et al., 2020), yet they suffer from limited interpretability (Lipton, 2018). According to the superposition hypothesis, dense vectors often compress sparse features into fewer dimensions, rendering them polysemantic (Arora et al., 2018; Elhage et al., 2022). This opacity creates a fundamental alignment gap (Luan et al., 2021), making it difficult to verify or intervene when retrieval processes diverge from human intent.

To bridge this gap, we propose applying Sparse Autoencoders (SAEs) (Ng et al., 2011) to the final output of sentence encoders, such as the E5 model (Wang et al., 2022). While SAEs are typically used to study internal LLM residual streams (Bricken et al., 2023; Cunningham et al., 2023), their potential to interpret dense retrievers remains underexplored. Specifically, we utilize a Top-k SAE architecture (Gao et al., 2024) to map dense embeddings into a higher-dimensional, sparse latent space. This methodology allows us to decompose entangled representations into monosemantic features that are human-interpretable, expanding the 1,024-dimensional input into a significantly larger latent space to resolve feature superposition.

In this work, we demonstrate that Top-k SAEs can effectively align dense sentence embeddings with human conceptual frameworks. Our experiments on the WikiText-103-v1 dataset show that the learned latent features correspond to granular concepts. Furthermore, we go beyond static analysis to demonstrate semantic steering. By manually clamping specific latent neurons, we show that it is possible to control the retrieval mechanism—filtering out unwanted concepts without retraining the backbone model. This capability addresses the limitations of macro-steering by enabling micro-level control over semantic features. This work suggests that SAE-based decomposition is a viable path toward transparent, aligned, and steerable neural search systems.

2 METHODOLOGY

We propose a framework to disentangle and align the dense representations of neural retrievers using a Top-k Sparse Autoencoder (SAE). This approach maps entangled dense embeddings into a high-dimensional sparse latent space where individual dimensions represent monosemantic concepts.

2.1 TOP-K SPARSE AUTOENCODER ARCHITECTURE

We adopt the Top-k SAE architecture to overcome the limitations of traditional L_1-regularized autoencoders. While L_1 penalties effectively induce sparsity, they often introduce a shrinkage bias that suppresses feature activation magnitudes, degrading reconstruction quality. (Bricken et al., 2023; Bussmann et al., 2024) The Top-k approach addresses this by directly enforcing a hard constraint (Makhzani & Frey, 2015): for every input embedding $\mathbf{x}$, only the k most active latent neurons are retained, while the rest are set to zero.

Our model projects the input embedding into a latent space $\mathbb{R}^{d_{latent}}$ (where $d_{latent} \gg d_{model}$), applies the Top-k activation function, and reconstructs the original vector from these sparse features. Following the approach of Gao et al. (2024), we tie the encoder and decoder weights to stabilize training. This ensures that the model learns a consistent and interpretable decomposition of the semantic space without the trade-offs inherent in soft sparsity constraints.

For the detailed mathematical formulation, and the auxiliary training objective used to mitigate dead neurons, please refer to Appendix B.

2.2 LATENT FEATURE STEERING VIA CLAMPING

A key contribution of our work is the ability to intervene in the retrieval process via latent feature clamping. Since each dimension i of $\mathbf{z}$ corresponds to an interpretable concept, we can surgically modify the latent vector to filter out unwanted semantic attributes from the retrieval results.

Given an input query $\mathbf{x}$, we compute its sparse representation $\mathbf{z}$. We define a clamping operation $C(\mathbf{z}, \mathcal{I}_{\text{clamp}})$ where $\mathcal{I}_{\text{clamp}}$ is the set of target neuron indices to deactivate:

$$\mathbf{z}_{\text{steered}}^{(i)} = \begin{cases} 0 & \text{if } i \in \mathcal{I}_{\text{clamp}} \\ \mathbf{z}^{(i)} & \text{otherwise} \end{cases} \tag{1}$$

By strictly zeroing out specific activations, this operation effectively removes corresponding concepts without affecting other semantic components. The steered embedding $\hat{\mathbf{x}}_{\text{steered}} = \mathbf{W}_{\text{dec}}\mathbf{z}_{\text{steered}}$ is then used for similarity search, enabling precise, inference-time control over the retrieval mechanism without fine-tuning the backbone model.

2.3 AUTOMATED FEATURE ANNOTATION PIPELINE

To scale the interpretation of thousands of latent features, we utilize an automated pipeline powered by GPT-4o-mini. The process follows three stages:

1. Orthogonality-based Filtering. We prioritize distinct features by calculating the Decoder Orthogonality (DO) for each neuron. Only neurons with a mean pairwise cosine similarity below a strict threshold are selected, ensuring that the annotated features are geometrically and semantically well-separated.

2. Context Extraction. For each target neuron, we retrieve the top-N sentences (e.g., $N = 10$) from the corpus that yield the highest activation values, providing a representative context of the learned pattern.

3. LLM Labeling with Coherence Check. An LLM acts as a linguist to identify common patterns across the retrieved sentences. Crucially, a Coherence Check is enforced: if the sentences lack a clear common

thread, the neuron is classified as *Unlabelable*. Otherwise, the LLM generates a specific **Label** and assigns a **Category** (e.g., Entity, Topic, Syntactic, or Pragmatic). Full details of the annotation prompts are provided in Appendix E.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Dataset & Model. We utilize the WikiText-103-v1 dataset (Merity et al., 2017) as our primary corpus. To ensure the quality of semantic representations, we applied a rigorous preprocessing pipeline involving sentence segmentation and length-based filtering, resulting in a final dataset of approximately 3.8 million sentences. Further details on the data preprocessing are provided in Appendix C.

For the backbone model, we employ `intfloat/e5-large-v2`, which generates 1,024-dimensional dense embeddings.

SAE Configuration. We train a Top-k SAE with a latent expansion factor of $12\times$, resulting in a latent dimension of $d_{latent} = 12,288$ given the input dimension of $d_{model} = 1,024$. The sparsity level is fixed at $k = 32$. To mitigate the dead neuron problem and ensure efficient feature utilization, we employ an auxiliary loss with $\alpha = 0.1$.

3.2 QUANTITATIVE ANALYSIS: DISENTANGLEMENT QUALITY

To verify that the SAE effectively disentangles the dense embedding space, we analyze two key metrics: **Decoder Orthogonality** and **Reconstruction Fidelity**. For detailed mathematical definitions of these metrics, please refer to Appendix D. Based on the trade-off between semantic separability and information retention, we selected the model with a latent dimension of $d_{latent} = 12,288$ ($12\times$ expansion) for our subsequent steering experiments.

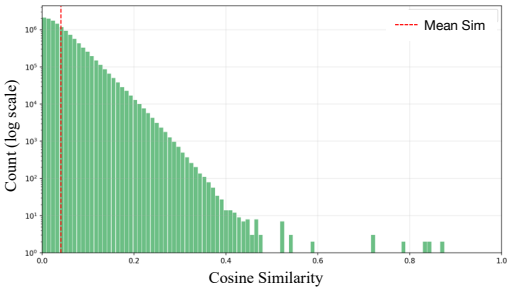

Figure 1: Neuron activation frequency distribution (log-scale) on the Wikipedia corpus.

Decoder Orthogonality (DO). A key prerequisite for interpretability is that learned features should represent distinct, non-overlapping concepts. We quantify this by measuring the mean pairwise cosine similarity between the columns of the decoder matrix $\mathbf{W}_{\text{dec}}$. As presented in Table 1, our $12\times$ model achieves a remarkably low mean orthogonality score of 0.0408. Figure 1 further illustrates the distribution of these pairwise similarities. The log-scale histogram reveals that the vast majority of feature pairs exhibit near-zero similarity, confirming that the learned semantic directions are structurally independent rather than redundant.

Model	Explained Variance	Mean DO	Dead Neurons (%)
$d_{latent} = 4,096$	0.9136	0.0371	0.49
$d_{latent} = 8,192$	0.9211	0.0389	1.01
$d_{latent} = 12,288$	0.9259	0.0408	0.74

Table 1: Reconstruction performance and sparsity metrics of Top-k SAEs. The expansion factor denotes the ratio of the latent dimension to the model dimension ($d_{model} = 1,024$). All models use $k = 32$.

Reconstruction Fidelity. Despite the aggressive sparsity constraint ($k = 32$), the model maintains high fidelity with an Explained Variance (EV) of 0.9259. This corresponds to a Fraction of Variance Unexplained (FVU) of less than 0.08, demonstrating that the sparse decomposition effectively preserves the essential semantic information of the original dense embeddings.

Feature Vitality and Sparsity. We analyze the activation frequency across the corpus to evaluate latent space utilization. As shown in Figure 2, the model exhibits a healthy long-tailed distribution with zero dead neurons (0.74% mortality). The median activation frequency of $\approx 10^{-5}$ confirms that features capture granular, rare contexts rather than generic noise, providing a robust foundation for subsequent steering experiments.

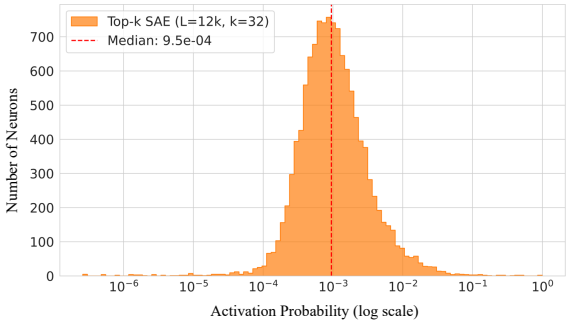

Figure 2: Neuron activation frequency distribution on the Wikipedia corpus. The characteristic long-tail indicates that the Top-k SAE effectively utilizes its expanded latent space to maintain feature sparsity and vitality.

3.3 Analysis of Auto-Labeled Features

We applied our automated pipeline to the trained Top-k SAE using GPT-4o-mini. Out of the neurons satisfying the orthogonality threshold, the LLM successfully generated coherent labels for the majority of features, confirming that the low-orthogonality criterion effectively filters for monosemantic concepts.

Granularity of Concepts. Unlike standard clustering which often stops at broad topics (e.g., "Politics" or "Arts"), the auto-labeled features exhibit remarkable granularity. For instance, the model captures specific abstract topics such as "Critical Reception of Media" (#7400) rather than general "Reviews," and distinguishes specific syntactic tokens like "Reference to Number 21" (#230) from general numeric values. As shown in Table 2, these features span across diverse categories including specific entities, pragmatic functions, and syntactic patterns.

ID	Category	Label	Representative Top Sentences
#11	Entity-Specific	*Rosetta & Rose Entities*	1) "Rosetta is the name of a lightweight dynamic translator..." 2) "Okafor attended Rosemont Elementary."
#7400	Topic-Specific	*Critical Reception of Media*	1) "Despite the overwhelming negative reviews, the film did receive some positive feedback." 2) "Overall, reception of the album was positive."
#487	Pragmatic	*Quotations and Citations*	1) "According to Mahatma Gandhi :" 2) "The Liber Eliensis described the situation as follows :"
#230	Syntactic/Token	*Reference to Number 21*	1) "His uniform number was 21." 2) "There were 21 fatalities."

Table 2: Selected examples of auto-labeled features representing diverse categories. The model captures highly granular concepts, ranging from specific entities to syntactic tokens and pragmatic functions. Each feature is illustrated with top-activating sentences.

3.4 Latent Steering for Alignment

Building on the interpretability confirmed above, we demonstrate the capability to align retrieval results with human intent via activation steering. By manually clamping specific latent neurons, we can surgically intervene in the retrieval process to modify semantic priorities without retraining the backbone model.

Consider a complex query: *"The **updated** **safety** **guidelines** applied only to heavy machinery operators."* We intervene by identifying and suppressing a specific semantic dimension:

- **Original Retrieval:** The baseline retrieval prioritizes the entangled context of safety and labor.
  - *This was largely done in **consideration of safety grounds** and usually applied to those conducting maintenance or the repair of equipment.*
  - *To ensure **safety** during the Stack period, the organizers maintained a perimeter around the working area, and allowed only **safety-trained** students through.*
- **Steering Intervention:** We identified neuron #4940 (*Industrial Safety*) via our auto-labeling pipeline and clamped its activation to zero ($z_{4940} = 0$).
- **Steered Retrieval:** Neutralizing the 'safety' component shifts results toward the regulatory scope; the original top result dropped to 9th place, while regulatory sentences ascended:
  - *The final rule adopted several **changes** to the HOS **regulations**, including a new **provision** requiring drivers to take a rest break during the work day under certain circumstances.*
  - *The policy has been **changed** so **permits** are only required for large scale film, video and photography requiring 10 person crews .*

These results demonstrate that Top-k SAEs effectively bridge dense vectors and human concepts. By modulating these semantic atoms, we align neural retrieval with specific intent, offering a principled framework for steerable, transparent search. See Appendix F for further case studies.

4 Conclusion and Future Work

In this work, we presented a framework for aligning the latent representations of dense retrievers with human conceptual structures utilizing Top-k Sparse Autoencoders. We demonstrated that applying the Top-k architecture allows for the successful decomposition of 1,024-dimensional embeddings into granular, monosemantic features, effectively bridging the gap between high-dimensional vector spaces and human interpretability.

Our automated analysis pipeline confirmed that these features correspond to specific entities, topics, and pragmatic functions. Crucially, we demonstrated that this disentanglement enables latent steering, which serves as a mechanism to intervene in the retrieval process and re-rank results according to user intent without model retraining.

Limitations. Our study focuses on a single backbone model (E5-large) and English benchmarks. Furthermore, we acknowledge that our hyperparameter selection is not exhaustive. While prior research indicates that scaling the latent dimension—often accompanied by increasing k—can improve performance up to the point of model collapse, we did not optimize for the ideal latent dimension or sparsity level. Since our primary objective was an exploratory investigation to verify if sentence embeddings could be decomposed into interpretable concepts and effectively steered, we maintained a fixed sparsity of $k = 32$. Consequently, the current performance may not represent the upper bound of the proposed framework.

Future Work. Our primary focus for future research is to integrate this steering mechanism into RAG pipelines to dynamically filter biased or unsafe concepts at inference time. We also plan to extend this methodology to multilingual settings to investigate cross-lingual conceptual alignment.

REFERENCES

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495, 2018.

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 Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders. *arXiv preprint arXiv:2412.06410*, 2024.

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.

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.

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. *arXiv preprint arXiv:2406.04093*, 2024.

Tibor Kiss and Jan Strunk. Unsupervised multilingual sentence boundary detection. *Computational linguistics*, 32(4):485–525, 2006.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

Edward Loper and Steven Bird. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*, 2002.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345, 2021.

Alireza Makhzani and Brendan J Frey. Winner-take-all autoencoders. *Advances in neural information processing systems*, 28, 2015.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017.

Andrew Ng et al. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19, 2011.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Vladimir Zaigrajew, Hubert Baniecki, and Przemyslaw Biecek. Interpreting CLIP with hierarchical sparse autoencoders. *arXiv preprint arXiv:2502.20578*, 2025.

DISCLOSURE OF AI USAGE

We acknowledge the use of Gemini-3 for linguistic polishing to improve the manuscript's readability. Furthermore, GPT-4o-mini was employed within our research pipeline to perform automated feature annotation (Section 3.3). The authors maintain full responsibility for the content and ensure that all AI-generated outputs were critically evaluated and verified by the authors.

A Qualitative Analysis of Learned Latent Features

In this section, we provide a detailed examination of the features learned by our Top-k Sparse Autoencoder. Table 3 presents a curated list of neurons, their human-interpreted labels, and representative activating sentences from the WikiText-103-v1 dataset. These examples demonstrate the model's ability to disentangle diverse linguistic and semantic concepts, ranging from specific entities to abstract topics and structural patterns.

ID	Label	Description	Representative Activating Sentences
#9689	Music Critical Reception	Summaries of professional reviews for songs or albums, typically citing "music critics".	1) "Upon its release, the track garnered **positive reviews from music critics**, who praised the song's composition..." 2) "The song received **generally favorable reviews from music critics** who commended Beyoncé's vocal performance..." 3) "The song received **generally mixed reviews from music critics**... both praising and criticizing the inclusion of the sample..."
#311	Disaster Impact	Descriptions of severe damage or "hardest hit" locations from storms/events.	1) "The **hardest hit areas** were in Jackson and Victoria counties where the **heaviest rains** fell." 2) "**Damage was heaviest** in the northern offshore islands and in the northern portion..." 3) "Coastal areas were **hardest hit**."
#11440	Entity: The Ramones	Specific references to the punk rock band 'The Ramones' and members.	1) "The **Ramones** were an American punk rock band that formed in... Forest Hills, Queens." 2) "**Ramones** is the debut studio album by the American punk rock band the **Ramones**..." 3) "All songs were written by the **Ramones**, except where noted."
#8870	Rhyme Scheme	Discussions of poetic structure, internal rhymes, and nursery rhymes.	1) "The **rhyme scheme** is AABB, or AA, B, CC, CB, B, B when accounting for internal **rhyme**." 2) "Most **rhyme schemes** are described using letters that correspond to sets of **rhymes**..." 3) "The poem's **rhyme scheme** is rhyming couplets rendered aa bb cc dd ee aa."
#1427	"Ajax" (Polysemantic)	Activates for AFC Ajax (Football), HMS Ajax (Ship), and Mythology.	1) "**Ajax** won the tie 4–1 on aggregate to progress to the second round." (Sport) 2) "The ship was renamed **Ajax** on 15 June 1869." (Naval) 3) "**Ajax** focuses on the proud hero of the Trojan War, Telamonian **Ajax**..." (Mythology)
#8408	Phrase: "There is"	Rhetorical or existential statements starting with "There is/are".	1) "**There has to be** something better." 2) "What else **is there**?" 3) "**There is** currently no authoritative voice classification system..."

Table 3: Selected interpretable features learned by the Top-k SAE ($d_{latent} = 12288$). We present three representative activating sentences per neuron to demonstrate semantic consistency.

B Detailed SAE Architecture and Training Objective

In this section, we provide the mathematical formulation of our Top-k Sparse Autoencoder and the training objectives employed.

B.1 Forward Pass Formulation

Let $\mathbf{x} \in \mathbb{R}^{d_{model}}$ be the dense sentence embedding generated by the backbone model. The encoding process consists of a linear transformation followed by a Top-k non-linearity:

$$\mathbf{z}_{\text{pre}} = \mathbf{W}_{\text{enc}}(\mathbf{x} - \mathbf{b}_{\text{dec}}) \tag{2}$$

$$\mathbf{z} = \text{TopK}(\text{ReLU}(\mathbf{z}_{\text{pre}}), k) \tag{3}$$

where $\mathbf{W}_{\text{enc}} \in \mathbb{R}^{d_{latent} \times d_{model}}$ is the encoder weight matrix, and $\mathbf{b}_{\text{dec}}$ is the decoder bias. The $\text{TopK}(\cdot, k)$ operation keeps only the k largest values of the pre-activations and zeroes out the rest, ensuring a fixed sparsity level.

The reconstruction $\hat{\mathbf{x}}$ is obtained via the decoder:

$$\hat{\mathbf{x}} = \mathbf{W}_{\text{dec}}\mathbf{z} + \mathbf{b}_{\text{dec}} \tag{4}$$

We enforce tied weights such that $\mathbf{W}_{\text{enc}} = \mathbf{W}_{\text{dec}}^T$, which acts as a regularization technique to improve training stability and feature coherence.

B.2 Training Objective with AuxK Loss

The primary objective is to minimize the reconstruction error, measured by the Mean Squared Error (MSE):

$$\mathcal{L}_{\text{recon}} = ||\mathbf{x} - \hat{\mathbf{x}}||_2^2 \tag{5}$$

A common challenge in training large-scale sparse autoencoders is the "dead neuron" problem, where certain latent features never activate across the entire dataset. To mitigate this, we incorporate an auxiliary loss term, $\mathcal{L}_{\text{aux}}$, often referred to as the AuxK loss. This objective encourages inactive neurons to predict the residual error $(\mathbf{x} - \hat{\mathbf{x}})$ of the main autoencoder, effectively "reviving" them by forcing them to explain the information missed by the currently active features.

The total loss function is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \alpha\mathcal{L}_{\text{aux}} \tag{6}$$

where α is a hyperparameter that controls the contribution of the auxiliary task. In our experiments, we set α to a non-zero value (e.g., 0.1) to proactively prevent feature collapse.

C Dataset Preprocessing Details

We constructed our training corpus using the WikiText-103-v1 dataset, which consists of over 100 million tokens extracted from verified "Good and Featured" articles on English Wikipedia. To adapt this document-level dataset for sentence embedding tasks, we performed the following preprocessing steps:

1. **Sentence Segmentation:** We split the raw articles into individual sentences using the NLTK sentence tokenizer (Kiss & Strunk, 2006; Loper & Bird, 2002) as a standard natural language processing tool.
2. **Filtering:** To remove noise and ensure semantic completeness, we applied several filters:

- We removed sentences shorter than 5 words to exclude section headers, list items, and fragmented text.
- We excluded extremely long sentences (e.g., > 128 words) to prevent excessive truncation by the backbone model's tokenizer.
- We filtered out non-natural language artifacts, such as URLs and wiki-markup boilerplate.

After this process, the final dataset consists of 3,837,611 unique sentences, which serve as the inputs for training our Top-k Sparse Autoencoders.

D Evaluation Metrics

Following the framework of Zaigrajew et al. (2025), we employ quantitative metrics to assess the quality of learned sparse representations.

D.1 Explained Variance (EV) and FVU

To measure reconstruction quality, we report the Explained Variance (EV). This metric is directly related to the Fraction of Variance Unexplained (FVU), where $\text{EV} = 1 - \text{FVU}$.

$$\text{EV} = 1 - \frac{\sum_{i=1}^{N} ||\mathbf{x}_i - \hat{\mathbf{x}}_i||_2^2}{\sum_{i=1}^{N} ||\mathbf{x}_i - \bar{\mathbf{x}}||_2^2} \tag{7}$$

A value close to 1.0 (or FVU close to 0.0) indicates that the sparse decomposition retains nearly all information present in the original dense embeddings.

D.2 Decoder Orthogonality (DO)

To evaluate the disentanglement of the learned features, we measure the orthogonality of the decoder weights. In an ideal disentangled representation, distinct latent features should correspond to distinct (orthogonal) directions in the semantic space.

Let $\mathbf{w}_i$ denote the i-th column vector of the decoder matrix $\mathbf{W}_{\text{dec}} \in \mathbb{R}^{d_{model} \times d_{latent}}$, representing the direction of the i-th latent feature. We first normalize all column vectors to unit length: $\hat{\mathbf{w}}_i = \frac{\mathbf{w}_i}{||\mathbf{w}_i||_2}$. We then compute the mean absolute pairwise cosine similarity between all unique pairs of features:

$$\mathcal{M}_{\text{ortho}} = \frac{1}{d_{latent}(d_{latent} - 1)} \sum_{i \neq j} |\langle \hat{\mathbf{w}}_i, \hat{\mathbf{w}}_j \rangle| \tag{8}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product. A lower $\mathcal{M}_{\text{ortho}}$ score indicates that the features are more independent and less redundant. A score of 0 implies perfect orthogonality among all feature directions.

E AUTOMATED LABELING PROMPT

To ensure transparency and reproducibility, we provide the full system and user prompts used for the automated feature annotation pipeline powered by GPT-4o-mini. The prompt was designed to enforce strict coherence checks and ensure specific, granular labels.

System Prompt

You are an expert linguist analyzing sentences that activate a specific SAE neuron. Your goal is to find a **specific, coherent pattern**.
CRITICAL INSTRUCTION: COHERENCE CHECK
Before labeling, check if the sentences share a strong semantic, syntactic, or pragmatic connection.

- If the sentences describe completely unrelated topics with no common thread, you MUST classify it as **'Unlabelable'**.
- Do NOT force a label if the pattern is weak or ambiguous.

If coherent, identify the **most specific** common thread possible (Entity, Issue, Pragmatic Role). All outputs must be in **English**.

User Prompt

Here are the sentences that most strongly activate Neuron #{neuron_idx}. The number in brackets is the activation score (higher is stronger).
[Sentences]
{List of top-20 activating sentences with scores}
Analyze the patterns deeply and output the result in JSON format with the following keys:

1. **"label"**:
   - If coherent: A **highly specific** label in English (3-4 words). e.g., "Opposition to Tax Reform".
   - If incoherent/mixed: **"Unlabelable"**
2. **"description"**: A detailed explanation of the pattern, including specific nuances, emotional tones, or context **in English** (1-2 sentences).
3. **"category"**: Choose one [Topic_Specific, Entity_Specific, Syntactic/Grammar, Pragmatic/Context, Polysemantic, Noise].
   - If 'Unlabelable', set this to **"Noise"** or **"Polysemantic"**.

Return ONLY the JSON string.

Figure 3: The prompts provided to the LLM for automated feature annotation.

F Additional Case Studies on Latent Feature Steering

To further validate the precision of semantic steering, we present additional case studies across various domains. Each case highlights how deactivating a single monosemantic neuron (identified via our automated pipeline) redirects the retrieval mechanism to prioritize alternative semantic facets of the query.

In the following tables, the values in brackets represent the cosine similarity scores between the query (original or steered) and the retrieved sentences.

F.1 Case Study 1: Neutralizing "Jumping Activities" (Neuron #2039)

The query *"The record for the longest jump was broken during the Olympic trials"* initially retrieves sentences focused exclusively on jumping events. By clamping neuron #2039, the focus shifts to the general concept of breaking records at Olympic trials across diverse sports.

Rank	**Original Retrieval (Top 5)**	**Steered Retrieval ($z_{2039} = 0$)**
1	[0.8483] ...record) and triple jump (44 feet 0).	[0.8548] In 1900, Olympian Maxie Long set the first official world record in the 400 meters...
2	[0.8449] Olympic skier Satre set a record jump of 112 feet (34 m) in 1937.	[0.8527] At the trials, she qualified for the finals... and broke the American record.
3	[0.8448] The record was eventually broken... who jumped 75 meters (246 ft).	[0.8521] ...set a new Olympic record, beating Phelps' previous record of 51.
4	[0.8445] ...Mike Powell, the world record holder in long jump.	[0.8495] ...Mike Powell, the world record holder in long jump.
5	[0.8436] ...eventual gold-medal-winning (and Olympic record) jump of 6 feet 4.	[0.8452] ...breaking the Olympic record with a time of 21. / surpassed the Olympic A standard.

Table 4: Retrieval results before and after deactivating the "Jumping Activities" feature.

F.2 Case Study 2: Neutralizing "Bridge Infrastructure" (Neuron #6188)

The query *"The bridge construction was delayed due to an update in safety requirements"* is suppressed by deactivating the bridge-specific context, shifting the results toward general construction delays and safety compliance.

Rank	**Original Retrieval (Top 5)**	**Steered Retrieval ($z_{6188} = 0$)**
1	[0.8647] The project was slated to have the 74-year-old bridge up to standards.	[0.8763] ...plans had to be revised to comply with new federal standards regarding steel pilings.
2	[0.8646] To remedy what was becoming a major delay... bridge to cross the shipping channel.	[0.8745] Construction was delayed again a month later, with work to begin in February 2016.
3	[0.8609] ...postponed until July 30, 2003, to improve safety on the highway.	[0.8710] Changes to the design and a lack of armor plating led to delays in building.
4	[0.8602] Construction was delayed again... for completion in April 2017.	[0.8705] It has been updated to include regulations on ship construction and safety.
5	[0.8601] ...plans had to be revised to comply with new federal standards.	[0.8691] But because the engineers needed to be re-certified, the start was delayed again.

Table 5: Retrieval results before and after deactivating the "Bridge Infrastructure" feature.

F.3 CASE STUDY 3: NEUTRALIZING "CLIMATE CHANGE AWARENESS" (NEURON #9054)

Suppression of the "Climate Change Awareness" feature in the query *"The government announced a major policy shift regarding carbon tax credits"* shifts focus from environmental activism to general fiscal and administrative policy changes.

Rank	**Original Retrieval (Top 5)**	**Steered Retrieval ($z_{9054} = 0$)**
1	[0.8524] The panel ultimately announced backing for a temporary carbon tax.	[0.8592] Its policy also represented a significant change from the idealism of previous governments.
2	[0.8460] ...revised act called for participation in international carbon markets.	[0.8574] This effectively represented a complete rewrite of UK energy policy for the future.
3	[0.8447] A carbon tax was introduced in 2012... but was scrapped in 2014.	[0.8545] ...the change was made because of privacy-related complaints.
4	[0.8434] This effectively represented a complete rewrite of UK energy policy.	[0.8521] Sweden launched a new 10-year environmental tax shift.
5	[0.8427] Abbott outlined his alternative climate change policy.	[0.8513] ...introduced a General Anti-Avoidance Rule to manage the risk of tax avoidance.

Table 6: Retrieval results before and after deactivating the "Climate Change Awareness" feature.