

# FEATURE HEDGING: CORRELATED FEATURES BREAK NARROW SPARSE AUTOENCODERS

David Chanin  
University College London

Tomáš Dulka  
Independent

Adrià Garriga-Alonso  
FAR AI

## ABSTRACT

It is assumed that sparse autoencoders (SAEs) decompose polysemantic activations into interpretable linear directions, as long as the activations are composed of sparse linear combinations of underlying features. However, we find that if an SAE is more narrow than the number of underlying “true features” on which it is trained, and there is correlation between features, the SAE will merge components of correlated features together, thus destroying monosemanticity. In LLM SAEs, these two conditions are almost certainly true. This phenomenon, which we call *feature hedging*, is caused by SAE reconstruction loss, and is more severe the narrower the SAE. In this work, we introduce the problem of feature hedging and study it both theoretically in toy models and empirically in SAEs trained on LLMs. We suspect that feature hedging may be one of the core reasons that SAEs consistently underperform supervised baselines. Finally, we use our understanding of feature hedging to propose an improved variant of matryoshka SAEs. Importantly, our work shows that SAE width is not a neutral hyperparameter: narrower SAEs suffer more from hedging than wider SAEs.

## 1 INTRODUCTION

As large language models (LLMs) are deployed in real-world applications, it is increasingly important to understand their internal workings. Sparse autoencoders (SAEs) decompose the dense, polysemantic activations of LLMs into interpretable latent features (Cunningham et al., 2024; Bricken et al., 2023) using sparse dictionary learning (Olshausen & Field, 1997). SAEs have the advantage of operating completely unsupervised, and can easily be scaled to millions of neurons in its hidden layer (hereafter called “latents”<sup>1</sup>) (Templeton et al., 2024; Gao et al., 2024).

While SAEs showed promising results, recent work has cast doubt on the performance of SAEs relative to baseline techniques. Wu et al. (2025) show that SAEs underperform on both concept steering and detection relative to baselines, and Kantamneni et al. (2025) show that SAEs underperform simple linear probes on both in-domain and out-of-domain detection, even when the probes have very few training samples. The question, then, is why do SAEs underperform relative to other techniques? And if we can identify the problems holding back SAEs, can we then fix those problems?

One fundamental issue with SAEs is the problem of feature absorption (Chanin et al., 2024), where a more specific latent suppresses the firing a more general latent. For instance, an SAE may have a latent that appears to track “Cities in USA” but that arbitrarily fails to fire on the specific cities “New York” and “Detroit”, where a city-specific latent fires instead. Feature absorption requires underlying features to exist in a hierarchy, with a parent feature  $f_p$  and a child feature  $f_c$ , where  $f_c$  can only fire if  $f_p$  is firing ( $f_c \implies f_p$ ). Feature absorption is caused by SAE sparsity penalty, and becomes more severe the wider the SAE. An SAE encoder/decoder under feature absorption is shown in Figure 1b.

In this paper, we identify another fundamental issue with SAEs which we call feature hedging. In hedging, an SAE is too narrow to represent both features  $f_a$  and  $f_b$  with their own latents  $l_a$  and  $l_b$ . Ideally, an SAE should assign a latent  $l$  to either  $f_a$  or  $f_b$ , and ignore the feature not being tracked. However, if  $f_a$  and  $f_b$  are either hierarchical as in absorption, or (anti-)correlated, then the SAE

<sup>1</sup>We use the term “latents” for the hidden neurons of the SAE to avoid overloading the term “feature”. We use “feature” only to describe interpretable concepts represented by the model.

Table 1: Comparing feature hedging and feature absorption

Feature absorption	Feature hedging
Learns gerrymandered latents	Learns polysemantic mixtures of features
Caused by sparsity loss	Caused by MSE reconstruction loss
Features are all tracked in the SAE	One feature is in the SAE, the other is not
Affects the encoder and decoder asymmetrically	Affects encoder and decoder symmetrically
Gets worse the wider the SAE	Gets worse the narrower the SAE
Requires hierarchical features	Requires only correlation between features

latent  $l$  can reduce reconstruction error by incorrectly mixing in components of both  $f_a$  and  $f_b$ . A sample SAE encoder and decoder experiencing hedging is shown in Figure 1a. In an LLM SAE, hedging will look like each SAE latent has noise mixed into it, reducing the performance of the latent for both detection and steering. Unlike with absorption, hedging becomes worse the narrower the SAE: thus trying to reduce absorption by making the SAE narrower will simply result in more hedging instead. The differences between hedging and absorption are shown in Table 1.

In LLM SAEs, the SAE is almost certainly narrower than the number of underlying features, as even extremely wide LLM SAEs appear to miss features (Templeton et al., 2024). Furthermore, we expect that nearly every feature in an LLM has positive and negative correlations to many features. We thus expect that hedging is the norm in LLM SAEs and will significantly distort their performance.

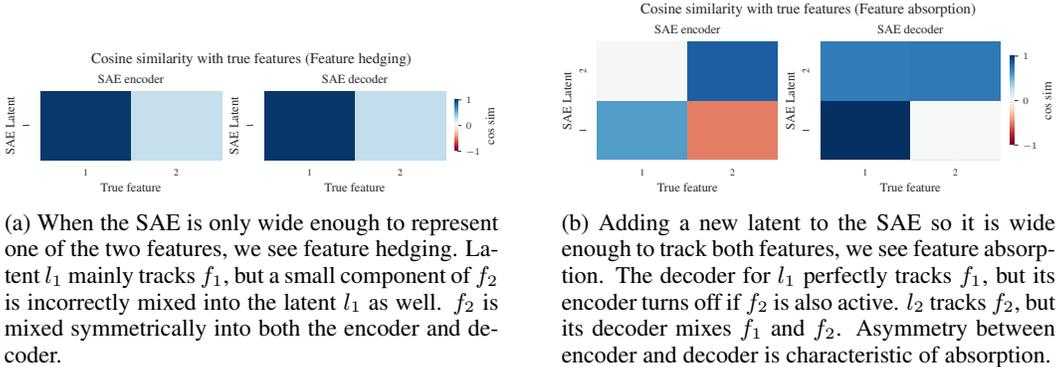


Figure 1: SAE encoder and decoder patterns for hierarchical features  $f_1$  and  $f_2$ , where  $f_1 \implies f_2$ . These features lead to either hedging or absorption depending on the width of the SAE.

A solution to feature absorption has been proposed in the form of matryoshka SAEs (Bussmann et al., 2025). Matryoshka SAEs use nested SAE loss terms to enforce a hierarchy on the SAE latents, solving absorption by forcing the narrow inner levels of the SAE to reconstruct inputs on their own. However, as we show in this paper, matryoshka SAEs suffer more from hedging due to the inner matryoshka levels essentially being very narrow SAEs. Matryoshka SAEs thus trade off absorption for hedging.

In this work, we define and study feature hedging both theoretically in toy models and empirically in LLM SAEs. We show that hedging is worse the more narrow the SAE, and introduce a technique to characterize the amount of hedging present in a given SAE. We also study hedging and absorption in matryoshka SAEs, and show that it is possible to improve the monosemanticity of matryoshka SAEs by tuning the relative loss coefficients in each level of the matryoshka SAE to better balance the competing forces of absorption and hedging—though both problems remain present. We show as well that SAE width is not a neutral hyperparameter: narrow SAEs suffer more from hedging than wider SAEs.

Code is available at <https://github.com/chanind/feature-hedging-paper>.

## 2 BACKGROUND

**Sparse autoencoders (SAEs).** An SAE decomposes an input activation  $x \in \mathbb{R}^D$  into a hidden state  $f$  consisting of  $L$  hidden neurons, called “latents”. An SAE is composed of an encoder  $W_{\text{enc}} \in \mathbb{R}^{L \times D}$ , a decoder  $W_{\text{dec}} \in \mathbb{R}^{D \times L}$ , a decoder bias  $b_{\text{dec}} \in \mathbb{R}^D$ , and encoder bias  $b_{\text{enc}} \in \mathbb{R}^L$ , and a nonlinearity  $\sigma$ , typically ReLU or a variant like JumpReLU (Rajamanoharan et al., 2024), TopK (Gao et al., 2024) or BatchTopK (Bussmann et al., 2024).

$$z = \sigma(W_{\text{enc}}(x - b_{\text{dec}}) + b_{\text{enc}}) \quad (1)$$

$$\hat{x} = W_{\text{dec}}z + b_{\text{dec}} \quad (2)$$

The SAE is trained with a reconstruction loss, typically Mean Squared Error (MSE), and a sparsity-inducing loss consisting of a function  $\mathcal{S}$  that penalizes non-sparse representation with corresponding sparsity coefficient  $\lambda$ . For standard L1 SAEs,  $\mathcal{S}$  is the L1 norm of  $f$ . For TopK and BatchTopK SAEs, there is no sparsity-inducing loss ( $\mathcal{S} = 0$ ) as the TopK function directly induces sparsity. There is sometimes also an additional auxiliary loss  $\mathcal{L}_{\text{aux}}$  with coefficient  $\alpha$  to ensure all latents fire. Standard L1 SAEs typically do not have an auxiliary loss (Olah et al., 2024). The general SAE loss is

$$\mathcal{L} = \|x - \hat{x}\|_2^2 + \lambda\mathcal{S} + \alpha\mathcal{L}_{\text{aux}}. \quad (3)$$

**Tied SAEs.** A tied SAE has  $W_{\text{enc}} = W_{\text{dec}}^T$ . The biases have different dimensions and are untied.

**Matryoshka SAEs.** A matryoshka SAE (Bussmann et al., 2025) extends the SAE definition by summing losses created by prefixes of SAE latents. This forces each sub-SAE to reconstruct input activations on its own, and incentivizes the SAE to place more common, general concepts into latents with smaller index number. A matryoshka SAE uses nested prefixes with sizes  $\mathcal{M} = m_1, m_2, \dots, m_n$  where  $m_1 < m_2 < \dots < m_n = L$ , where  $L$  is the number of latents in the full dictionary. Matryoshka SAE loss is:

$$\mathcal{L} = \sum_{m \in \mathcal{M}} (\|x - \hat{x}_m\|_2^2 + \lambda\mathcal{S}_m) + \alpha\mathcal{L}_{\text{aux}} \quad (4)$$

Where  $\hat{x}_m$  is the reconstruction for the SAE using the first  $m$  latents, and  $\mathcal{S}_m$  is the sparsity penalty applied to the first  $m$  latents. For TopK and BatchTopK Matryoshka SAEs, there is no sparsity penalty ( $\mathcal{S}_m = 0$ ) as the TopK function directly imposes sparsity.

## 3 TOY MODELS OF FEATURE HEDGING

The linear representation hypothesis (LRH) states that features in LLMs are represented as nearly-orthogonal linear directions in representation space (Bricken et al., 2023). The goal of SAEs, then, is to recover these underlying “true features” of the model, where each latent of the SAE decoder perfectly matches an underlying feature of the model. In real LLMs we do not have ground-truth knowledge of these underlying features, making it difficult to know if SAEs are succeeding at recovering model features. Fortunately, it is easy to create synthetic training data for SAEs that follows the LRH and gives us ground-truth knowledge of the underlying features. This allows us to understand when SAEs will learn the underlying features of the model, and when SAEs fail.

We define a toy model consisting of  $N$  true features  $F \in \mathbb{R}^{N \times D}$ , where each  $\|f_i\|_2 = 1$ . These features are mutually orthogonal, so  $\forall i \neq j, f_i \cdot f_j = 0$ . Each feature  $f_i$  has a corresponding firing probability  $p_i \in [0, 1]$ . For each sample, we generate a binary activation vector  $\mathbf{a} \in \{0, 1\}^N$  where  $a_i \sim \text{Bernoulli}(p_i)$  indicates whether feature  $f_i$  is active (fires). The model can incorporate feature dependencies by conditioning firing probabilities:  $a_i | \mathbf{a}_{-i} \sim \text{Bernoulli}(p_i(\mathbf{a}_{-i}))$ , where  $\mathbf{a}_{-i}$  denotes the activation states of all other features. We then generate SAE training samples  $x$  from this model as  $x = \sum_{i=1}^N a_i f_i$ .

We say that an SAE is *correct* or *monosemantic* for this toy model if every latent in the SAE dictionary matches a true feature direction, and each SAE latent corresponds to a different true feature. Formally, there exists a bijection  $\pi : \{1, \dots, L\} \rightarrow \{1, \dots, N\}$  such that  $\cos(W_{\text{dec},i}, f_{\pi(i)}) = 1$  for all  $i \in \{1, \dots, L\}$ . We only investigate SAEs where  $L \leq N$  in our toy experiments. We say an SAE is *polysemantic* if some SAE latents contain positive or negative components of multiple true features, so there exists at least one latent  $i \in \{1, \dots, L\}$  such that  $|\{j \in \{1, \dots, N\} : |W_{\text{dec},i} \cdot f_j| > \epsilon\}| > 1$  for some threshold  $\epsilon > 0$ .

For all SAEs in this section, we train on 15M synthetic activations using SAELens (Bloom et al., 2024). In this section we show plots of the cosine similarity between the SAE encoder / decoder and the true features. Each cell  $i, j$  in these plots is simply  $\cos(W_{\text{enc},i}^T, f_j)$  and  $\cos(W_{\text{dec},i}, f_j)$ , respectively. We re-arrange the indices of the SAE latents to best align visually with true features.

### 3.1 FULLY INDEPENDENT FEATURES

We first study the case of a toy model with  $N = 4$  independent features. Features 1-3 fire with probability 0.25, and feature 4 fires with probability 0.2. We plot the encoder / decoder cosine similarity with true features in Figure 2. When features fire independently, the SAE learns correct features regardless of the width of the SAE.

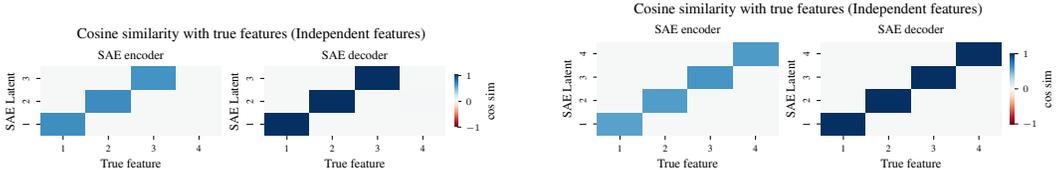


Figure 2: SAE with three latents (left) and four latents (right) trained on a toy model with independent features. Both SAEs learn correct features.

Unfortunately, real LLMs do not have fully independent features. SAEs were first studied under toy models with independent features (Elhage et al., 2022), and this is likely why the field was not aware of feature hedging earlier.

### 3.2 HIERARCHICAL FEATURES

Next, we explore true features that fire hierarchically. We modify the toy model from Section 3.1 above, and set  $f_3$  as the parent feature and  $f_4$  as the child feature, so  $f_4 \implies f_3$ . That is,  $f_4$  cannot fire unless  $f_3$  is also firing. Hierarchical features cause feature absorption in SAEs that are wide enough to represent both the parent and child feature, but what happens if the SAE is not wide enough to represent the child latent? This is important as this is the intuition behind why Matryoshka SAEs work to combat absorption: if inner SAE levels are too narrow to represent both parent and child features, we hope that only the parent will be represented. We show results in Figure 3.

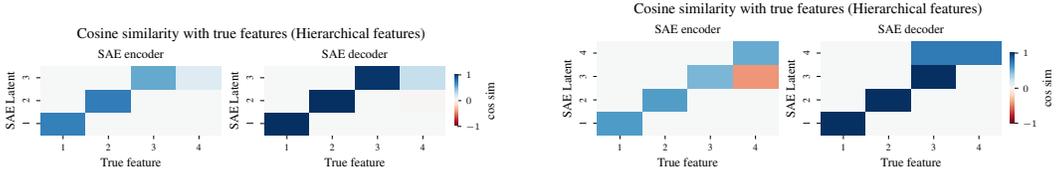


Figure 3: SAEs trained on a toy model with hierarchical features ( $f_3 \implies f_4$ ). When the SAE is too narrow to represent  $f_4$  (left), we see hedging where latent 3 mixes . When the SAE is wide enough to contain both  $f_3$  and  $f_4$  (right), we see feature absorption.

As expected, in the full-width SAE we see a classic feature absorption pattern. The parent latent encoder,  $l_3$ , learns  $\neg f_4 \wedge f_3$ , disabling the latent from firing if  $f_4$  is present. The child latent,  $l_4$ , mixes both  $f_3$  and  $f_4$  together in the decoder.

However, when the SAE is too narrow to represent  $f_4$ , we now see that latent  $l_3$  mixes a component of  $f_4$  into both its encoder and decoder! We refer to this as *feature hedging*. The SAE is learning an incorrect, polysemantic latent that mixes correlated features.

While we expect this will be a problem for all SAEs, it is particularly problematic for Matryoshka SAEs, as Matryoshka SAEs combat absorption by using inner SAE levels that are too narrow to contain both parent and child features. However, as we see here, this causes hedging.

We show that MSE loss directly causes hedging with hierarchical features in Appendix A.1.

### 3.3 POSITIVELY CORRELATED FEATURES

Hierarchy is a particularly extreme form of positive correlation, where a feature can only fire if another feature also fires. Next, we relax that restriction, and investigate what happens if a feature is merely more likely to fire along with another feature, but can still fire on its own as well. We modify the toy model so that  $p_4 = 0.2$  if  $f_3$  fires, but  $p_4 = 0.1$  if  $f_3$  does not fire, adding a small positive correlation between  $f_3$  and  $f_4$ . We show results Figure 4

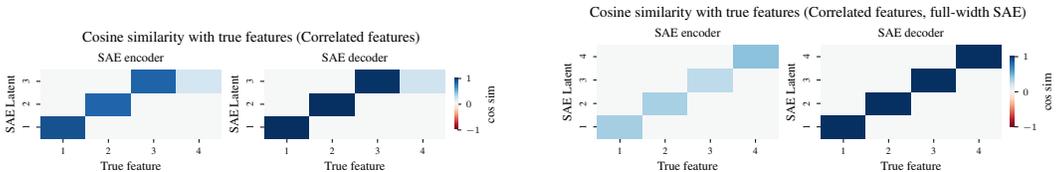


Figure 4: SAEs trained on a toy model with positive correlation between features  $f_3$  and  $f_4$ . When the SAE is too narrow to represent  $f_4$  (left), we still see hedging in latent 3. When the SAE is wide enough to contain both  $f_3$  and  $f_4$  (right), the SAE learns correct features.

We still see that the SAE is mixing a positive component of  $f_4$  into  $l_3$ , despite there no longer being perfect hierarchy! We also see that if we extend the SAE width so that  $f_4$  is tracked by its own latent  $l_4$ , there is now no absorption at all, as absorption requires (nearly) hierarchical features to arise.

### 3.4 ANTI-CORRELATED FEATURES

So far we have only seen the effect of positive correlation between features. We next change our toy model so  $f_4$  is more likely to fire if  $f_3$  does *not* fire. We set  $p_4 = 0.1$  if  $f_3$  fires, but  $p_4 = 0.2$  if  $f_3$  does not fire. Results are shown in Figure 5.

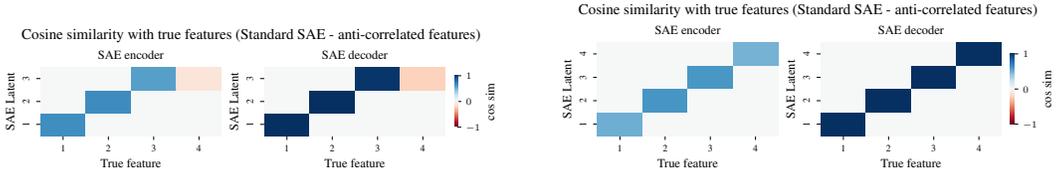


Figure 5: SAEs trained on a toy model with negative (anti) correlation between features  $f_3$  and  $f_4$ . When the SAE is too narrow to represent  $f_4$  (left), we latent 3 mixes a *negative* component of  $f_4$ . When the SAE is wide enough to contain both  $f_3$  and  $f_4$  (right), the SAE learns correct features.

We now see that  $l_4$  is mixing in a *negative* component of  $f_3$ . This demonstrates that the correlation is the cause of the hedging: flipping the sign of the correlation flips the sign of the hedging.

The implications of this for SAE performance are quite dire. While it is already bad for positively correlated features to become hedged (e.g. “sunshine” and “summertime”), at least the mixed features have some relation to each other. For anti-correlated features, this could look like a latent for “chemical molecules” having a negative component of the “Darth Vader” feature mixed in, since chemical modules and Darth Vader are highly anti-correlated. Worse, it is not even clear if the inverse of the “Darth Vader” direction is meaningful in the model at all, or is just noise. We further

expect that there are many more negative correlations than positive correlations in language, e.g. a negative component every word in every non-English language may be mixed into every latent tracking an English word. These negative correlations likely introduce what looks like a lot of random noise into SAE latents, and this can only harm performance and interpretability.

### 3.5 STUDYING HEDGING IN SINGLE-LATENT SAES

We next investigate hedging in the simplest possible toy SAE setting: an SAE with a single latent. We use a model with two true features  $f_1$  and  $f_2$  ( $N = 2, D = 50$ ). Each feature fires with magnitude 1.0. Unless otherwise specified,  $f_1$  fires with probability 0.25, and  $f_2$  fires with probability 0.2. We use SAELens (Bloom et al., 2024) to train a single-latent SAE on these activations.

#### 3.5.1 FULLY INDEPENDENT FEATURES

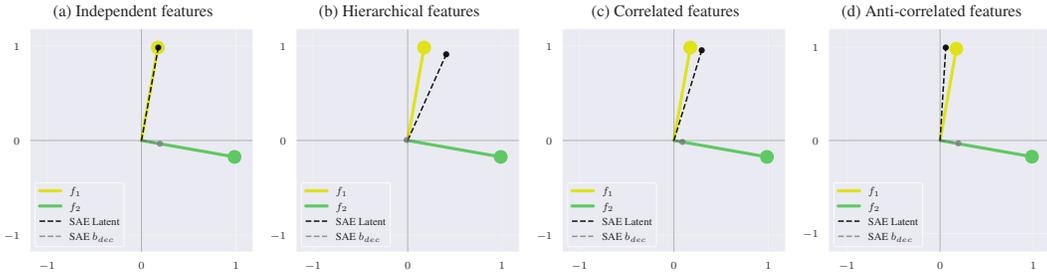


Figure 6: True features, and SAE decoder latent and  $b_{dec}$  for single-latent SAE and a toy model with two true features. When the features fire independently, there is no hedging seen in the SAE latent. When any correlation is present, the SAE latent shows clear hedging.

We first study the case when  $f_1$  and  $f_2$  fire independently. We find that the SAE correctly represents  $f_1$  without any interference from  $f_2$ . However, the decoder bias has incorrectly learned to represent the direction of  $f_2$ , but with magnitude 0.2, equal to the probability of  $f_2$  firing. The single SAE latent, SAE bias term, and true features are shown in Figure 6a.

We consistently find this pattern of the decoder bias merging in positive components of features not tracked by their own latent. In this sense, the decoder bias can be thought of as an always-on latent, and thus is thus also susceptible to hedging.

#### 3.5.2 HIERARCHICAL FEATURES

Next, we investigate what happens if  $f_1$  and  $f_2$  are in a hierarchy, so  $f_2$  can only fire if  $f_1$  fires, but  $f_1$  can still fire on its own ( $f_2 \implies f_1$ ). We adjust the firing probability of  $f_2$  so that  $P(f_2|f_1) = 0.2$ , and  $P(f_2|\neg f_1) = 0$  (thus,  $P(f_2) = 0.05$ ). In a two-latent SAE this setup would cause feature absorption. We plot the SAE latent, decoder bias, and true features in Figure 6b.

Here we clearly see feature hedging. The single SAE latent has now merged in a component of  $f_2$  into its single latent, so it is now a mixture of  $f_1$  and  $f_2$ . This merging of features reduces the MSE loss of the SAE despite being a degenerate solution.

#### 3.5.3 POSITIVELY CORRELATED FEATURES

Next, we change our setup so that  $P(f_2|\neg f_1) = 0.1$  instead of 0. We still keep  $P(f_2|f_1) = 0.2$ , so that  $f_2$  is more likely to fire if  $f_1$  fires, but it can still fire on its own as well. The features are now merely correlated rather than following a strict hierarchy. Results are shown in Figure 6c.

We still see hedging in the SAE latent, but less than with full hierarchical features. However, if the L1 penalty is high enough and the level of correlation is low enough, then the SAE can still learn the correct features, as positive hedging increases the L0 of the SAE slightly relative to learning just  $f_1$ . We show the resulting SAE latent and features with high L1 penalty in Figure 7b. Interestingly, we now see that the hedging has moved more apparently into the decoder bias instead. If we use a full-width SAE, the SAE learns the true features despite the correlation (Appendix A.2).

### 3.5.4 ANTI-CORRELATED FEATURES

Next, we reverse the conditional probabilities of  $f_2$  so that  $P(f_2|f_1) = 0.1$  and  $P(f_2|\neg f_1) = 0.2$ . Now  $f_2$  is more likely to fire on its own than it is to fire along with  $f_1$ , making these features anti-correlated. Results are shown in Figure 6d.

Now the SAE latent has actually merged a *negative* component of  $f_2$  into its single latent instead of a positive component. How does this work? We see that the decoder bias,  $b_{dec}$ , has a larger component of  $f_2$  than in the positive correlation case. The SAE is using the decoder bias to include a “default” value for  $f_2$ , and then when  $f_1$  fires, the SAE latent’s negative component of  $f_2$  acts to reduce the amount of  $f_2$  present in the reconstruction. The SAE is abusing the correlation to adjust its guess of the amount of  $f_2$  that should be output despite not having a dedicated latent for  $f_2$ : if  $f_1$  is active, then the likelihood that  $f_2$  is active decreases, and the SAE likewise reduces the amount of  $f_2$  that is output.

Increasing L1 penalty cannot solve this, as the negative component of hedging in the encoder does not increase L0 of the SAE. If we use a full-width SAE, we again see the SAE learns the true features despite the correlation (see Appendix A.2).

### 3.5.5 HEDGING IS A FUNCTION OF FEATURE CORRELATION

Next, we explore the effect of feature correlation on the amount of hedging in our single-latent, two feature setting. We set  $P(f_1) = 0.45$  and  $P(f_2) = 0.25$ , but change the correlation between these features,  $\rho$ , to range from  $-0.5$  to  $0.5$ . We then calculate the cosine similarity of the SAE decoder latent,  $l$ , with  $f_2$ . We furthermore initialize the single SAE latent to match  $f_1$ , so that any deviation from this must be caused by gradient pressure rather than simply being an unfortunate local minimum. If there is no hedging occurring, then  $\cos(l, f_2) = 0$ , as we saw in Figure 6a. Results are shown in Figure 7a.

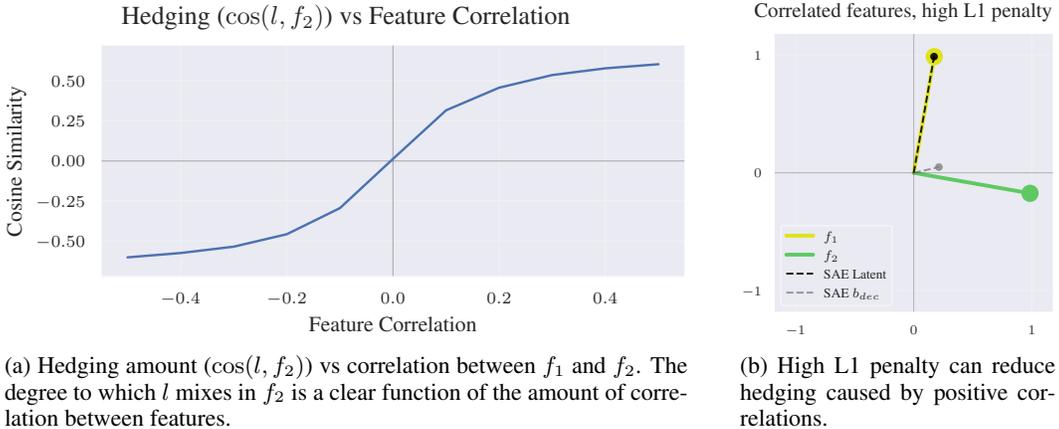


Figure 7: Hedging as a function of feature correlation, and effect of L1 penalty on positive hedging.

As expected, the amount of hedging directly tracks the amount of correlation. The hedging also matches the sign of the correlation as well, with negative correlation resulting in a negative component of  $f_2$  being mixed into  $l$ , and positive correlation resulting in a positive component of  $f_2$  being mixed into  $l$ .

## 4 QUANTIFYING HEDGING IN LLM SAES

While we have demonstrated hedging in a synthetic setting, it remains a question how much hedging occurs in LLM SAES. Based on our understanding of hedging in toy models, we expect that when a new latent is added to an SAE, this should “pull out” the component of the new feature from existing SAE latents, where it was previously hedged. Thus if hedging occurs, *the change in existing latents after a new latent is added should project onto that new latent*. If hedging did not exist, then adding a new latent should not have any effect on existing latents.

**Parent latents are learned before child latents** A key assumption in Matryoshka SAEs is that if latents exist in a hierarchy, and the SAE is too narrow to represent both the parent and child, the SAE will learn the parent first. We feel this assumption is reasonable since parent latents, by definition, fire more frequently than child latents, so the SAE is incentivized to learn them first.

This insight allows us to differentiate hedging from absorption. Under absorption, if a newly added latent is a child feature of an existing latent, then the encoder for the parent latent adds a negative component of the child latent to avoid firing when the child is active, but *the parent decoder latent remains unchanged*. This corresponds to adding  $l_2$  to Figure 1a and arriving at Figure 1b. The decoder of  $l_1$  (the parent) remains identical to before  $l_2$  is added, except the *hedging from  $f_2$  is removed*. Thus, changes to existing decoder latents cannot be absorption and must be due to hedging.

**Hedging degree** Taking this into account, we define a metric called hedging degree,  $h$ . We take an existing SAE  $s_0$  with  $L$  latents and add  $N$  new latents to the SAE. After adding these latents, we continue training the SAE and arrive at a new SAE,  $s_1$ , with  $L + N$  latents. We also continue training  $s_0$  on the same tokens that we train  $s_1$  on to ensure that any difference between  $s_0$  and  $s_1$  is due only to the newly added latents.  $W_{\text{dec}}^0$  refers to the new decoder of  $s_0$ , and  $W_{\text{dec}}^1$  refers to the decoder of  $s_1$ .  $W_{\text{dec}}$  is normalized so each latent has unit norm. We define the difference in the original  $L$  latents between  $s_0$  and  $s_1$  as:

$$\delta_L = W_{\text{dec}}^1[0 : L] - W_{\text{dec}}^0[0 : L] \tag{5}$$

where  $W_{\text{dec}}^1[L : L + N]$  refers to the newly added decoder latents.  $W_{\text{rand}}[0 : N]$  refers to a decoder consisting of  $N$  randomly initialized unit-norm latents. All decoders are normalized to have latents of unit norm. We define the projection of a vector  $v$  onto a subspace spanned by  $W$  as:

$$\text{Proj}(v, W) = \|W(W^T W)^{-1} W^T v\| \tag{6}$$

We expect that even if there were no hedging at all, simply due to noise, existing SAE decoder latents may undergo a change that has some small projection onto new added latents. We want to make sure that anything we quantify as hedging must be larger than what we would expect from random noise. Taking this into account, the hedging degree  $h$  is then defined as:

$$h = \frac{1}{L} \sum_i^L \underbrace{\|\text{Proj}(\delta_L[i], W_{\text{dec}}^1[L : L + N])\|}_{\text{Projection of } \delta_L \text{ onto } N \text{ new latents}} - \underbrace{\|\text{Proj}(\delta_L[i], W_{\text{rand}}[0 : N])\|}_{\text{Projection of } \delta_L \text{ onto } N \text{ random latents}} \tag{7}$$

Any value of  $h > 0$  corresponds to hedging above what we would expect from random noise, as  $h$  subtracts the projection along  $N$  randomly initialized unit-norm latents as part of the computation.

The choice of the number of new latents  $N$  is a hyperparameter of hedging degree. We use  $N = 64$  for our hedging degree calculation. We explore the effect of different choices on  $N$  in Appendix A.4.

#### 4.1 RESULTS

We experiment with SAEs trained on Gemma-2-2b (Team et al., 2024), as this model is commonly used for SAE research due to the thoroughness of the Gemma Scope suite of SAEs (Lieberum et al., 2024), as well as Llama-3.2-1b (Dubey et al., 2024) to validate results on another LLM. All SAEs are trained first on 250M tokens of the Pile uncopyrighted (Gao et al., 2020). After adding  $N = 64$  latents, we continue training for another 250M tokens. The version of the SAE without latents added is also trained for another 250M tokens, so each SAE is trained for 500M tokens total. The pair of extended and non-extended SAEs is used to calculate hedging degree. SAE training details are in Appendix A.3.

We first calculate hedging degree vs SAE width in Figure 8a, with widths ranging from 128 to 65536. Hedging degree is dramatically higher at narrower widths, especially at 4096 width and below. While the hedging rate drops a lot with increasing SAE width, even at our max width of 65536 no SAE achieves 0 hedging degree, indicating there is still hedging occurring.

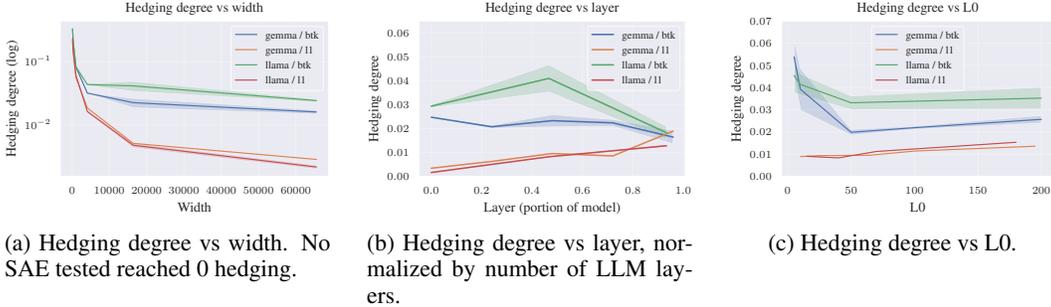


Figure 8: Hedging degree for SAEs trained on Gemma-2-2b layer 12. Unless otherwise specified, SAEs have width 8192, BatchTopK SAEs have K=25. Shaded area in plots is 1 std.

We next calculate hedging degree vs L0 (the average number of active latents) in Figure 8c, with L0 ranging from about 5 to 200. Very low L0 seems to lead to more hedging for BatchTopK SAEs, but the effect is minor compared with the effect of SAE width on hedging degree.

Finally, we calculate hedging degree vs layer in Figure 8b. The hedging degree for L1 and TopK SAEs appears to merge around the end of the SAE, but overall the layer does not appear to have a massive effect on hedging degree.

It also appears that BatchTopK SAEs have more hedging than L1 SAEs. We suspect that L1 loss can reduce hedging from positively correlated features, as we saw in Section 3.5.3.

We further validate hedging in LLM SAEs via a case-study of adding a new latent to an SAE trained on Gemma-2-2b in Appendix A.5.

## 5 BALANCING HEDGING AND ABSORPTION IN MATRYOSHKA SAEs

Matryoshka SAEs (Bussmann et al., 2025) combat absorption with nested SAE loss prefixes. Each level acts like a small SAE, and is forced to reconstruct the input on its own. This forces the SAE to learn more general concepts in earlier levels, and makes it difficult for the SAE to make holes in the recall of parent latents for absorption, as this would hurt the reconstruction of earlier levels.

However, since early matryoshka levels are effectively narrow SAEs, they suffer from feature hedging. As we saw in Section 4.1, the more narrow an SAE is, the worse the hedging. Matryoshka SAEs thus solve feature absorption at the expense of exacerbating feature hedging.

Inspecting the effect of hedging and absorption on the SAE encoder in Figure 1b shows that hedging and absorption have opposite effects. For hierarchical features, hedging adds a positive component of child features into the parent encoder latent, but absorption does the opposite and adds a negative component of child features into the parent latent. If we balance the negative component of child latents from absorption with the positive component from hedging, these effects can cancel out.

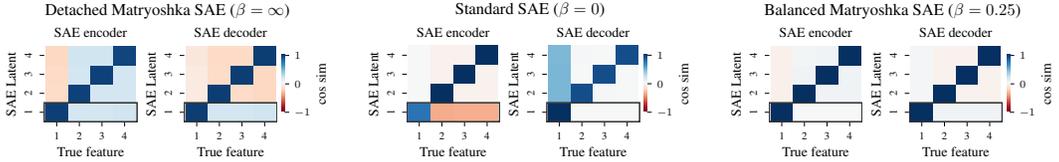
**Balance matryoshka SAE** We extend the definition of a matryoshka SAE from Equation 4 to allow applying a scaling coefficient  $\beta_m$  to the loss for each matryoshka level:

$$\mathcal{L} = \sum_{m \in \mathcal{M}} \beta_m (\|a - \hat{a}_m\|_2^2 + \lambda \mathcal{S}_m) + \alpha \mathcal{L}_{\text{aux}} \tag{8}$$

We refer to this extension as a *balance matryoshka SAE*, where each  $\beta_m \geq 0$  controls the relative balance of each level. If each  $\beta_m = 1$  this is a standard matryoshka SAE. If  $\beta_m = 0$  for all matryoshka levels except the outer-most level, this reduces to a standard (non-matryoshka) SAE.

We demonstrate this balancing in a toy model of hierarchical features. The toy model has 4 features, with feature 1 being the parent feature and features 2-4 being children (features 2-4 can only fire if feature 1 is also firing). Feature 1 fires with probability 0.25, and each child feature fires with probability 0.15 if feature 1 is firing. We train a matryoshka SAE with a single inner level consisting

of only latent 1 with balance coefficient  $\beta$  (Since there is only one inner level, we always set the outer level coefficient to 1). For more details on this toy setup, see Appendix A.7.

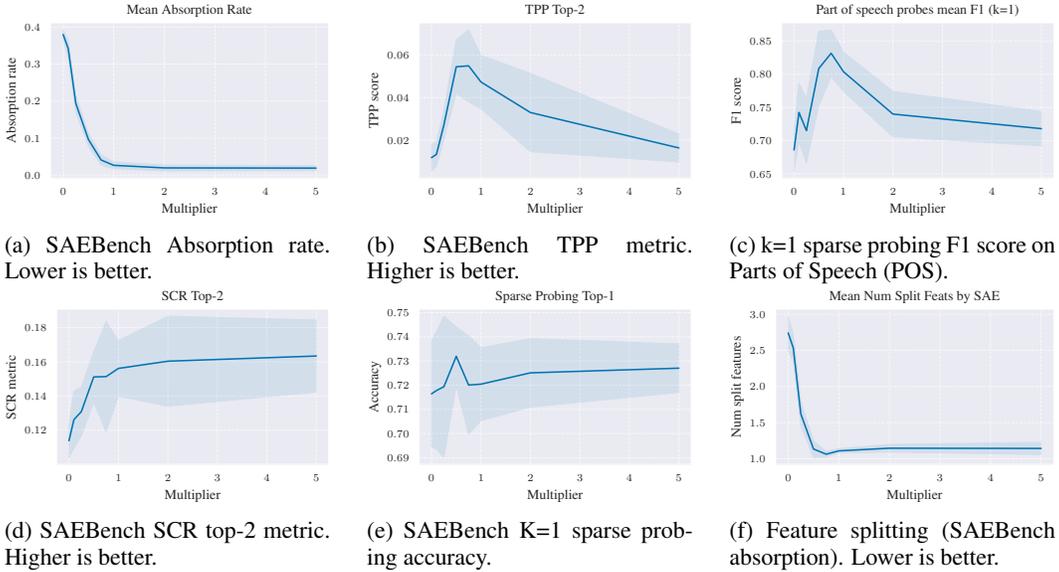


(a) Matryoshka SAE with detached loss (equivalent to a matryoshka SAE with  $\beta = \infty$ ). Hedging adds positive components of the child features 2-4 to the encoder of latent 1. (b) Standard SAE (equivalent a matryoshka SAE with  $\beta = 0$ ). Absorption adds negative components of the child features 2-4 to the encoder of latent 1. (c) Roughly balanced matryoshka SAE with  $\beta = 0.25$ . The positive and negative contributions hedging and absorption roughly cancel out, leaving a nearly perfect SAE.

Figure 9: Balancing hedging and absorption in a toy model of hierarchical features. Child features 2-4 only fire if parent feature 1 fires. The matryoshka SAE has a single inner level with 1 latent, represented by a black box around latent 1.

We show results in Figure 9. When  $\beta$  is too high or too low this results in hedging or absorption, respectively. When  $\beta = 0.25$ , these balance out and the SAE learns a near perfect representation.

Next, we train LLM balance matryoshka SAEs with different balance ratios on Gemma-2-2b layer 12. The SAEs are BatchTopK with  $k=40$ , trained on 500M tokens. The SAEs have 5 matryoshka levels of sizes 128, 512, 2048, 8192, and 32768 (so the full SAE has width 32768). We set the outermost  $\beta_5 = 1$ , and set a constant multiplier between each subsequent  $\beta_m$ , so multiplier =  $\beta_m / \beta_{m+1}$ . If the multiplier is 0.5, then  $\beta_m = 0.5^{(5-m)}$ .



(a) SAEBench Absorption rate. Lower is better. (b) SAEBench TPP metric. Higher is better. (c)  $k=1$  sparse probing F1 score on Parts of Speech (POS). (d) SAEBench SCR top-2 metric. Higher is better. (e) SAEBench  $K=1$  sparse probing accuracy. (f) Feature splitting (SAEBench absorption). Lower is better.

Figure 10: Performance of balance matryoshka SAEs vs multiplier. The shaded area is 1 std. Multiplier=0 is equivalent to a standard SAE, and multiplier=1 is a standard matryoshka SAE.

We train 10 seeds for each multiplier and show results in Figure 10 for absorption rate, targeted probe perturbation (TPP), Spurious Concept Removal (SCR),  $K$ -sparse probing, and feature-splitting metrics from SAEBench (Karvonen et al., 2025), and  $k=1$  sparse probing results (Gurnee et al., 2023) for a Parts of Speech (POS) dataset we created using Treebank POS tagged sentences (Marcus et al., 1993). We add a POS dataset for probing since POS are very general concepts, and should be learned in the earliest levels of a matryoshka SAE.

For TPP, feature splitting, and sparse probing, using a compound multiplier of around 0.75 achieves better results than either a standard matryoshka SAE or a standard (non-matryoshka) SAE, providing evidence that balancing matryoshka losses can improve the performance. Using a multiplier of 0.75 still scores well on the absorption metric as well. Strangely, SCR appears to perform better at higher multipliers. However, SCR is also the noisiest metric, and the noise is higher at high multipliers, so it could be that hedging increases the noise of the SCR metric but does not fully break it. We provide further results and more details in Appendix A.9.

While balancing each  $\beta_m$  can improve performance on most metrics, we do not expect this to perfectly solve absorption and hedging. We show in Appendix A.8 that balancing all hedging and absorption with a single  $\beta_m$  is not always possible. We expect it may be possible to further improve performance by learning different balancing coefficients per latent, but this is left to future work.

## 6 RELATED WORK

Other work has highlighted theoretical problems with SAEs. Till (2024) investigated a problem where SAEs may increase sparsity by inventing features. For instance, an SAE may fabricate a “red triangle” feature in addition to “red” and “triangle” features. Templeton et al. (2024) discuss the problem of feature splitting, where an SAE may not learn features at a desired level of specificity. Engels et al. (2024) investigates SAE errors and finds that SAE error may be pathological and non-linear. Engels et al. (2025) further shows that there are features that cannot be expressed as a simple linear direction, and thus SAEs may struggle to represent these features. Wu et al. (2025) and Kantamneni et al. (2025) both investigate the empirical performance of SAEs and find that SAEs underperform baselines.

## 7 DISCUSSION

SAEs remain a promising technique for decomposing the residual stream of LLMs in an unsupervised manner. However, given recent work showing that SAEs underperform relative to baselines (Wu et al., 2025; Kantamneni et al., 2025), it is imperative that we understand the reasons for this underperformance so they can be addressed.

In this work, we introduced the problem of feature hedging in SAEs, showing it both theoretically in toy models, and empirically in SAEs trained on real LLMs. We suspect that hedging, along with absorption, may be one of the core theoretical problems leading to poor SAE performance.

Using our understanding of hedging, we introduced the balance matryoshka SAE architecture, allowing balancing of hedging and absorption against each other, improving interpretability. We view balance matryoshka SAEs as a starting point, and expect this architecture can be improved by optimizing the balance coefficients. There may not be a single coefficient that perfectly balances hedging and absorption for all features, so we expect there may be further gains from learning a different balancing coefficients per latent in the SAE. We leave these improvements to future work.

### ACKNOWLEDGMENTS

This work was carried out as part of the ML Alignment & Theory Scholars (MATS) program. DC was supported thanks to EPSRC EP/S021566/1. We are grateful to Andrew Mack, Juan Gil, Satvik Golechha, Josh Engels and Henning Bartsch for discussions and input during the project.

### REFERENCES

- Joseph Bloom, Curt Tigges, Anthony Duong, and David Chanin. Saelens. <https://github.com/jbloomAus/SAELens>, 2024.
- Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermy, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2, 2023.
- Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders. *arXiv preprint arXiv:2412.06410*, 2024.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. *arXiv preprint arXiv:2503.17547*, 2025.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *arXiv preprint arXiv:2409.14507*, 2024.

Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=F76bWRSLeK>.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Manan Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuwei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li,

Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keenally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vitor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaoqian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.

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.

Joshua Engels, Logan Riggs, and Max Tegmark. Decomposing the dark matter of sparse autoencoders. *arXiv preprint arXiv:2410.14670*, 2024.

Joshua Engels, Eric J Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=d63a4AM4hb>.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

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.

- Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL <https://openreview.net/forum?id=JYs1R9IMJr>.
- Subhash Kantamneni, Joshua Engels, Senthooan Rajamanoharan, Max Tegmark, and Neel Nanda. Are sparse autoencoders useful? a case study in sparse probing. *arXiv preprint arXiv:2502.16681*, 2025.
- Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. Saebench: A comprehensive benchmark for sparse autoencoders in language model interpretability, 2025. URL <https://arxiv.org/abs/2503.09532>.
- Tom Lieberum, Senthooan Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2, August 2024.
- Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.
- Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=I4e82CIDxv>.
- Chris Olah, Adly Templeton, Trenton Bricken, and Adam Jermyn. April update. <https://transformer-circuits.pub/2024/april-update/index.html>, 2024. URL <https://transformer-circuits.pub/2024/april-update/index.html>.
- Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.
- Senthooan Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhu-patiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshtir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin,

Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Rostrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Faret, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL <https://arxiv.org/abs/2408.00118>.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Summers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. <https://transformer-circuits.pub/2024/scaling-monosemanticity/>, May 2024. Accessed on May 21, 2024.

Demian Till. Do sparse autoencoders find true features? *LessWrong*, 2024. URL <https://www.lesswrong.com/posts/QoR8noAB3Mp2KBA4B/do-sparse-autoencoders-find-true-features>.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Axbench: Steering llms? even simple baselines outperform sparse autoencoders. *arXiv preprint arXiv:2501.17148*, 2025.

## A APPENDIX

### A.1 HEDGING IS CAUSED BY RECONSTRUCTION LOSS: CURVES FOR SINGLE-LATENT SAEs

What causes hedging? We hypothesize that it is a combination of not enough latents to represent every feature, and the fact that MSE loss incentivizes reconstructing multiple features imperfectly as opposed to only one feature perfectly.

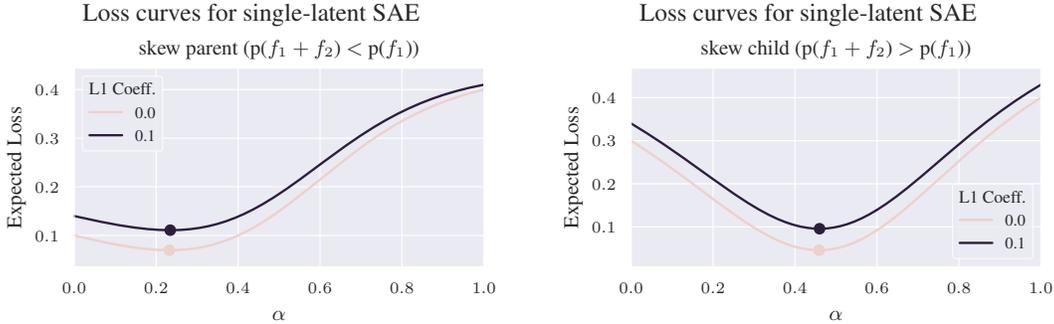
To test this, we analyze the loss curves for a single-latent tied SAE with a parent-child relationship between the two features  $f_1$  and  $f_2$ , so  $f_2 \implies f_1$ . The ideal SAE latent must be some combination of these two features. As there are no other interfering features to break the symmetry between encoder and decoder, the SAE can be expressed by a single unit norm latent. We set the SAE latent  $l$  to an interpolation of these two features,  $l = \alpha f_2 + (1 - \alpha)f_1$  (adjusted to have unit norm). We calculate expected SAE loss consisting of MSE + L1 loss for  $0 \leq \alpha \leq 1$ .

First, we set  $P(a = f_1) = 0.3$  and  $P(a = f_1 + f_2) = 0.1$ . We characterize the probabilities this way since there are only two firing possibilities we need to consider: either  $f_1$  is firing on its own or  $f_1$  and  $f_2$  are firing together. We use L1 coefficient of 0 and 0.1 to explore the effect of the sparsity penalty on loss. We also consider the case where both features fire together more than they fire on their own, with  $P(a = f_1) = 0.1$  and  $P(a = f_1 + f_2) = 0.3$ . Loss curves are shown in Figure 11.

In these plots,  $\alpha = 0$  corresponds to the SAE latent being exactly  $f_1$ , and  $\alpha = 1$  corresponds to the latent being  $f_2$ , and  $\alpha = 0.5$  corresponds to  $f_1 + f_2$ . We clearly see that the SAE loss has a single minimum between  $f_1$  and  $f_1 + f_2$ , showing that the MSE minimum is attained with feature hedging.

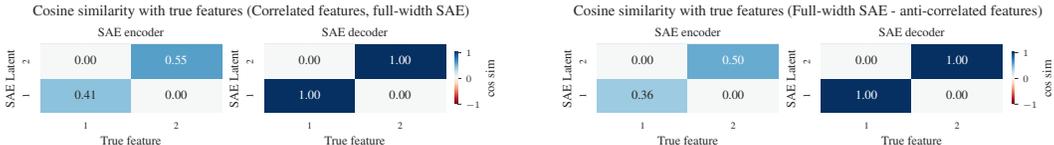
### A.2 FULL-WIDTH SAE TOY MODEL RESULTS

We extend the discussion of single-latent SAEs to explore what happens if the SAE has two latents, the same number of latents as the number of true features. We use the same toy model as in Section 3.5.3 for the positive correlation case, and the same toy model as in Section 3.5.4 for the anti-correlated case. We use L1 penalty of  $1e-3$  for the positive correlation case, the same as the L1 penalty that caused hedging in single-latent SAEs.



(a) Loss curves when the parent feature  $f_1$  fires more on its own than with child feature  $f_2$ . Loss is minimized between  $f_1$  and  $f_2$  rather than at  $f_1$  ( $\alpha = 0$ ). Sparsity penalty does not change the minimum.  
 (b) Loss curves when the parent feature  $f_1$  fires less on its own than it does with the child feature  $f_2$ . Loss is incorrectly minimized between  $f_1$  and  $f_2$ . Sparsity penalty does not change the minimum.

Figure 11: Loss curves for an SAE with a single latent  $l$  and 2 hierarchical features, where  $f_2 \implies f_1$ . The minimum loss is indicated with a dot on each plot.  $\alpha = 0$  means that  $l = f_1$ , and  $\alpha = 1$  means  $l = f_2$ . In all cases, loss is minimized when the latent  $l$  is a combination of  $f_1$  and  $f_2$ .



(a) Full-width SAE with correlated features. The SAE is still able to perfectly learn the underlying features despite the correlation.  
 (b) Full-width SAE with anti-correlated features. The SAE is still able to perfectly learn the underlying features despite the correlation.

Figure 12: Full-width SAE results on correlated and anti-correlated toy models.

We plot the results in Figure 12. In both cases, the full-width SAEs are able to perfectly recover the true features despite the correlation, and despite the low L1 penalty. This shows that hedging is caused by the SAE being too narrow, as increasing the width of the SAE solves the problem.

### A.3 TRAINING DETAILS FOR LLM SAES

All SAEs are trained on the Pile uncopyrighted (Gao et al., 2020), using a batch size of 4096 activations and context length of 1024 tokens. For L1 SAEs, we use a linear L1 warm-up of 10k steps. SAEs are trained on a single 80gb Nvidia H100 GPU. Model weights are loaded in fp32 precisions, but autocast to bfloat16 during training. We initialize the SAE so that the encoder and decoder are identical, where each latent has norm 0.1, following the procedure described in (Olah et al., 2024). All L1 SAEs are trained with learning rate  $7e-5$ , and BatchTopK SAEs are trained with learning rate  $3e-4$ . SAEs are trained using SAELens (Bloom et al., 2024).

Unless otherwise specified, BatchTopK SAEs use  $k=25$ . For SAEs trained on Gemma-2-2b, we conduct most experiments at layer 12 (roughly in the middle), and L1 SAEs trained on Gemma-2-2b use L1 coefficient of 10. This coefficient does not result in dead extension latents, and yields a L0 around 50. For SAEs trained on Llama-3.2-1b, we conduct most experiments at layer 7 (roughly in the middle of the model), and for L1 SAEs trained on Llama-3.2-1b, we use L1 coefficient of 0.5. This coefficient does not result in dead extension latents, and yields a L0 around 50.

### A.4 CHOICE OF HEDGING HYPERPARAMETER N

Our hedging degree metric requires adding  $N$  new latents onto an existing SAE to extend it, naturally leading to the question of what is a reasonable choice of  $N$ . We plot hedging degree vs  $N$  for Gemma-2-2b layer 12, given an initial BatchTopK SAE of width 8192 in Figure 13. We find that hedging degree increases until about  $N=250$ . We choose  $N=64$  for our experiments, as 64 is still a small

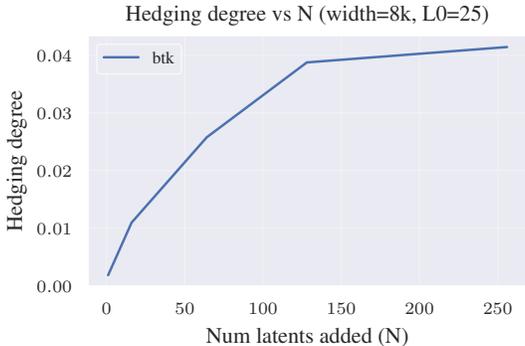


Figure 13: Hedging degree vs N

number of latents relative to the size of the residual stream (2304 for Gemma-2-2b), while still being large enough to hopefully reduce noise from any specific latent that gets added. Furthermore, as we see in the plot, the hedging degree from N=64 is about in the middle of the curve, further validating that this is a reasonable choice.

#### A.4.1 EXTENDING LLM SAES

We train two versions of extension SAES - one for L1 loss SAES and one for BatchTopK SAES. In both cases, we begin with a pretrained SAE and add  $N$  latents randomly initialized with norm 0.1, and with the same encoder and decoder directions, following Olah et al. (2024). For the BatchTopK SAES, we simply train the SAE from this point as normal, as the TopK auxiliary loss (Gao et al., 2024) will naturally ensure that the newly added latents do not simply die off.

For L1 SAES with high L1 penalty, dead latents become a more serious problem. We find that most of the newly added extension latents will rapidly be killed off if we simply train as normal. To combat this, we re-warm-up the L1 penalty. However, we cap the minimum L1 penalty at  $\lambda_{\min}$ , so for the portion of the warm-up where the L1 penalty would normally be below  $\lambda_{\min}$ , the L1 penalty is left at  $\lambda_{\min}$  instead. This capping helps ensure the existing SAE latents are not very disturbed by this change in the L1 penalty. If the final L1 penalty is  $\lambda_{\min}$  or below, then we do not perform this warm-up at all, as the L1 penalty is not strong enough to immediately kill off the newly added latents.

For Gemma-2-2b SAES, we set  $\lambda_{\min} = 10.0$ . For Llama-3.2-1b SAES, we set  $\lambda_{\min} = 0.5$ .

This warm-up procedure is only used for the high-L1 variants in Figure 8c - for all other plots the L1 coefficient used is less than  $\lambda_{\min}$ , so no warmup is needed.

#### A.5 CASE STUDY: ADDING A NEW LATENT TO AN EXISTING SAE

We next explore how hedging affects a real SAE. We trained a L1 SAE on Gemma-2-2b layer 12 with width 8192 for 250M tokens on the Pile (Gao et al., 2020), then add a new latent to the SAE, and continue training both the original SAE and the extended SAE for another 250M tokens.

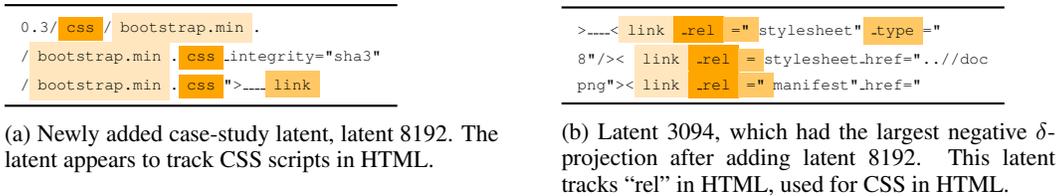


Figure 14: Sample top activating examples for case study latents.

We examine inputs that cause the newly added latent to fire to get a sense of what it represents. We reproduce a portion of the top activating examples for the new latent in Figure 14a. This latent appears to fire on CSS scripts included in HTML. A larger set of inputs is shown in Appendix A.6.

Next, we look at the magnitude of change in existing latents projected on the new latent. Based on our understanding of hedging, if a latent loses a large component of the newly added latent, this corresponds to a likely hierarchical relationship with the new latent. The latent which lost the largest component of the new latent is latent 3094, which seems to track the “rel” HTML attribute used mainly for linking CSS scripts. We show top activating examples for latent 3094 in Figure 14b.

Since CSS scripts are just one type of asset that can be linked using “rel”, this appears to be exactly the sort of hierarchical relationship we expect to be heavily impacted by hedging.

### A.6 ADDITIONAL CASE STUDY DASHBOARDS



Figure 15: Dashboard for the newly added case study latent representing CSS scripts in HTML.



Figure 16: Dashboard for latent 3094, representing the “rel” HTML attribute used for CSS scripts. This latent has the highest negative  $\delta$ -projection on the newly added case study latent.

### A.7 TOY BALANCE MATRYOSHKSA SAEs

To explore the effect of balancing matryoshka losses in a simple toy setting, we create a toy model with 4 true features, all mutually orthogonal and with unit norm in a 50 dimensional space. We set up a hierarchical relationship between these features, so feature 1 fires with probability 0.25, and features 2, 3, and 4 all fire with probability 0.15 only if feature 1 fires. Thus, feature 1 is the parent feature in the hierarchy and features 2, 3, and 4 are all child features.

We train a matryoshka SAE with 4 latents on 100,000,000 samples from this toy model. The matryoshka SAE has a single inner level consisting of 1 latent, to match the number of parent latents in our hierarchy. Since our goal with this toy is just to build intuition, we initialize the SAE to the correct solution and allow the training to thus pull it away from this correct solution. This also

ensures that each variation of our SAE with different balancing co-efficients learns the same latents in the same order, so visual comparison is easy.

### A.8 TOY UNBALANCEABLE MATRYOSHKA SAEs

The situation above where each child feature has the same probability of firing is unrealistic - we would expect that child features all fire with different probabilities from each other. Can we still balance the SAE perfectly in this situation? We adjust the toy model from above so that the 3 child features fire with probabilities 0.02, 0.2, and 0.5 for features  $f_2$ ,  $f_3$ , and  $f_4$ , respectively. We then try to manually balance this SAE, finding that  $\beta = 0.17$  gives roughly the best balance. We plot the resulting encoder/decoder cosine similarities in Figure 17.

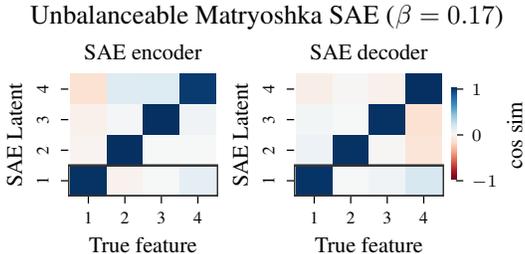


Figure 17: SAE encoder and decoder vs true feature cosine similarities for a balance matryoshka SAE where the child features fire with different probabilities. It’s no longer possible to perfectly balance all 3 child features with the same  $\beta$ , but we can still do reasonably well.

We now see it is no longer possible to choose a single  $\beta$  that perfectly balances all 3 children. We see slight hedging of feature 4 in latent 1, and slight absorption of feature 2 in latent 1. Still, this looks decent compared to the full hedging or full absorption scenario, so we still expect that while balancing is not a perfect solution, it should be an improvement. We believe it should be possible to finding ways of better balancing the contribution of each outer latent on each inner latent, but this is left to future work.

### A.9 SAE EVALUATION

#### A.9.1 SAE BENCH EVALS

We evaluate our SAEs mainly using SAE BENCH (Karvonen et al., 2025). All evals are performed using default settings. We run all evaluations on an Nvidia H100 GPU with 80gb GPU memory. We evaluate on the following SAE BENCH tasks:

**K-sparse probing** k-sparse probing builds on the work of Gurnee et al. (2023), where the goal is to create a linear probe from model activations using only  $k$  neurons as input to the probe. This was adapted for use as an SAE evaluation technique by Gao et al. (2024). We focus mainly on  $k = 1$  sparse probing, which means finding the single best SAE latent that serves as a classifier for a given concept, and evaluating the accuracy of that latent. SAE BENCH includes supervised classification datasets on which k-sparse probing is evaluated.

**Feature absorption** The feature absorption metric in SAE BENCH is a variation on the metric defined in the original feature absorption work (Chanin et al., 2024). This metric uses a first-letter spelling task and first identifies the “main” latents for that task using k-sparse probing (Gurnee et al., 2023). Then, the metric identifies cases where a linear probe is able to correctly perform the first-letter classification task, but the “main” SAE latents fail to perform the task. The metric also looks for other latents that project onto the linear probe direction and fire in place of the “main” latents. Lower absorption is better.

The SAE BENCH absorption metric also includes “absorptions fraction”, “feature splitting”, and “first-letter k=1 sparse probing” results as well, which we include in our extended results. Absorption fraction detects partial absorption, where a parent latent can still fire but weaker when an absorbing

child latent fires as well. Feature splitting detects the amount of interpretable feature splitting occurring in the SAE. Interpretable feature splitting is still considered negative, as we would prefer that features do not split at all and the SAE can still represent general, high-level concepts. The k-sparse probing results for the first-letter spelling task is calculated as part of the absorption metric, but is an interesting sparse-probing result in and of itself.

**Spurious concept removal (SCR)** SCR builds on the SHIFT method from Marks et al. (2025) to detect how well an SAE isolates concepts. The metric uses datasets where two properties are perfectly entangled, for instance “profession” and “gender”, and trains a biased probe on these concepts. The SCR metric then detects how well  $k$  SAE latents can be ablated to de-bias the probe. If the SAE latents learn disentangled concepts, then it should only take a few SAE latents to perfectly de-bias the probe. A high SCR score means the SAE latents represent disentangled concepts.

**Targeted probe perturbation (TPP)** The TPP metric extends SCR to multi-class labels. Binary probes are trained for each class, and TPP measures how well ablating  $k$  SAE latents can degrade the performance of just one of the probes without degrading performance on the other probes. A high TPP score means that concepts are represented by distinct sets of SAE latents, rather than latents being entangled with many concepts.

#### A.9.2 PARTS OF SPEECH (POS) PROBING DATASET

We are interested as well in general, high-frequency concepts that we expect should be learned in the inner-most levels of a matryoshka SAE. These concepts should be the most affected by both absorption and hedging, as these concepts can be considered parent concepts to most other concepts. Parts of speech (POS) is a great test-case for these general concepts, and are not covered by the SAE Bench sparse probing task. As such, we create our own custom POS dataset using the Penn Treebank tagged sentences (Marcus et al., 1993).

We simplify the Treebank parts of speech to the following core set:

"TO", "IN", "DT", "CC", "NNS", "PRP", "POS"

We pass these tagged sentences through an LLM, and then collect activations for the final token of position of each word at a given layer in the LLM. We create a binary classification dataset for each of these parts of speech, and perform k-sparse probing (Gurnee et al., 2023) on SAE latents to find the top  $k$  latents that act as a classifier for each of these parts of speech.

#### A.9.3 BALANCE MATRYOSHKA SAE FULL RESULTS

#### A.10 LIMITATIONS

We only test hedging in SAEs up to 65k latents on LLMs with 2b parameters due to compute constraints. Our method for detecting hedging requires fine-tuning SAEs, which is expensive.

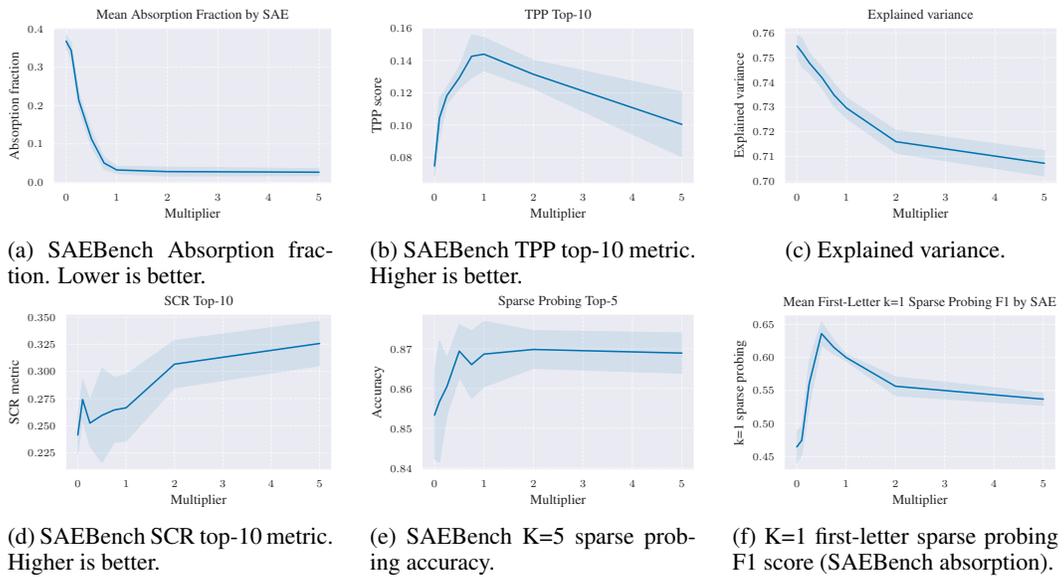


Figure 18: Performance of balance matryoshka SAEs vs multiplier for extended metrics. The shaded area in the plots refers to 1 std. Multiplier=0 is equivalent to a standard non-matryoska SAE, and multiplier=1 is equivalent to a standard matryoshka SAE.