Table 1: Sample max activating examples for latents 7112 and 7657 for Gemma Scope 16k, layer 0, 105 L0 from Neuronpedia. The token where the SAE latent activates is highlighted in yellow. Latent 7112 appears to be a lowercase "L" starting-letter latent, and latent 7657 appears to be a corresponding uppercase "L" latent.

LATENT 7112	LATENT 7657
žda se naplaćuje naknada	LC, an aluminum boat
. E. Søli, 20	as LIFT and LF-Net. Once
a></code> </ul	latter's sister Louise, who in

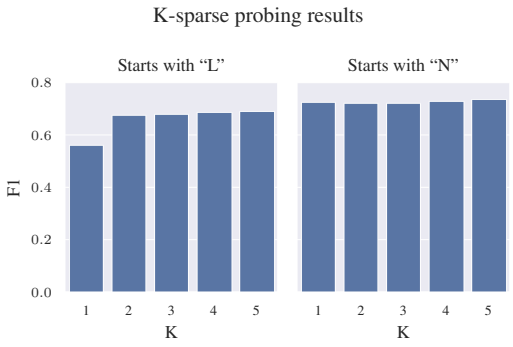
5.3 Measuring feature splitting and feature absorption

Feature splitting A key phenomenon identified from previous studies of SAEs is feature-splitting [2], where a feature represented in a single latent in a smaller SAE can split into two or more latents in a larger SAE. During our experiments, we found strong evidence of feature-splitting in the Gemma Scope SAEs.

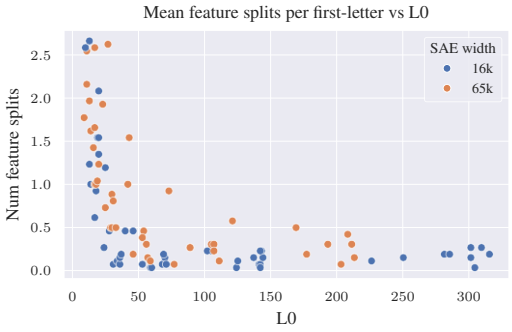
For instance, in the layer 0, 16k width, 105 L0 SAE, we find two encoder latents (id:7112 and id:7657 [1]) which align with the "L" starting letter probe. Inspecting max activating examples, we see latent 7112 activates on tokens starting with lowercase "l", while 7657 activates on tokens starting with uppercase "L". Some activating examples for these latents are shown in Table 1.

Feature splitting like this is not necessarily problematic for interpretability efforts since the split features are still easily identifiable, and depending on the context it may be more useful to have either a single "starts with L" latent or a pair of "starts with uppercase / lowercase L" latents.

We measure feature splitting using k-sparse probing [12] on SAE activations. If increasing the k-sparse probe from k to $k+1$ causes a significant increase in probe F1 score, then the additional SAE latent provides a meaningful signal, and the combination of these $k+1$ latents is likely a feature split. In the example of the uppercase "L" and lowercase "l" split, a k-sparse probe with $k=2$ trained on both these latents should predict "starts with letter L" much better than either latent on its own. Figure 8a shows F1 vs K for letters "L" and "N". The "L" k-sparse probe shows a significant jump in F1 score moving from k=1 to k=2 corresponding to feature splitting, while the F1 score for the "N" k-sparse probe is relatively constant.



(a) K-sparse probing results for letters "L" and "N", layer 0, 16k width, 105 L0. "L" shows a significant improvement in F1 between k=1 and k=2 corresponding to feature splitting.



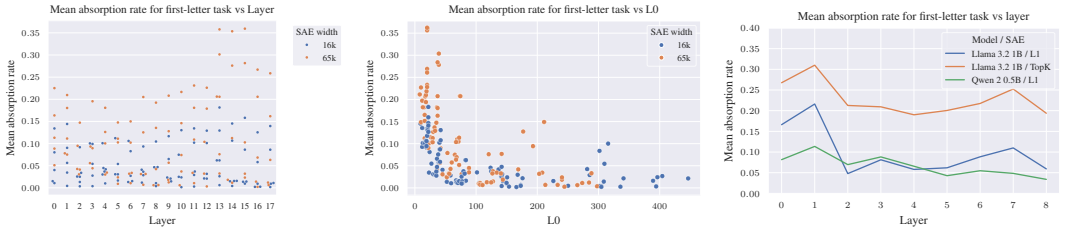
(b) Mean number of feature splits per letter on the first-letter spelling task by L0. Feature splitting occurs more frequently with higher sparsity.

Figure 8: Feature splitting analysis in sparse autoencoders

We detect feature splitting by measuring whether increasing k by one causes a jump in F1 score by more than a threshold, τ. We use $\tau = 0.03$ as a reasonable choice after inspecting situations like in Figure 8a, where feature splitting corresponds to an F1 score jump between 0.05 - 0.1. Figure 8b shows feature splitting vs L0 for all 16k and 65k width Gemma Scope SAEs.

[1] `https://www.neuronpedia.org/list/cm0h1n2mt00019jdk274owq9e`

The single latent or a set of traditional feature split latents that seem to act as a classifier for a human-interpretable feature like "starts with S" fail to fire in a seemingly arbitrary number of cases. What fires instead are often approximately token-aligned latents with small but positive alignment with the LR probe. We say these latents are absorbing the feature.



(a) Mean feature absorption rate vs layer on the first-letter task, Gemma Scope 16k and 65k SAEs. We do not see an obvious pattern in absorption rates by layer.

(b) Mean feature absorption rate vs L0 on first-letter task, Gemma Scope 16k and 65k SAEs. Wider and more sparse SAEs demonstrate higher rates of absorption.

(c) Mean feature absorption rate vs layer on the first-letter task on Llama 3.2 1B and Qwen 2 0.5B models, L1 loss and TopK SAE architectures, layers 0-8.

Figure 9: Feature absorption rates

We quantify the extent to which feature absorption occurs with the metric **feature absorption rate**. We first find k feature splits for a first-letter feature using a k-sparse probe. We then find false-negative tokens that all k feature-split SAE latents fail to activate on, but which the LR probe correctly classifies, and run an integrated-gradients ablation experiment on those tokens. The ablation effect finds the most causally important SAE latents for the spelling of that token. If the SAE latent receiving the largest negative magnitude ablation effect has a cosine similarity with the LR probe above 0.025, and is at least 1.0 larger than the latent with the second highest ablation effect, we say that feature absorption has occurred. These thresholds were chosen from manual inspection of the data to best distinguish the absorption phenomenon. We then calculate feature absorption rate as:

$$\texttt{absorption\_rate} = \frac{\texttt{num\_absorptions}}{\texttt{lr\_probe\_true\_positives}}$$

If there are more than 200 false negatives per letter, we randomly pick 200 samples to estimate the number of absorptions. We see absorption rate increases with higher sparsity and higher SAE width. Lower L0 likely pushes the SAE to absorb dense features like spelling information, increasing feature sparsity. Feature absorption rate vs L0 for Gemma Scope SAEs layers 0-17 is shown in Figure 9b. Absorption rate by letter is shown in Appendix A.14. We also train our own set of standard L1 loss SAEs on the first 8 layers of Qwen2 0.5B [32] and Llama 3.2 1B [6], and TopK SAEs [10] on Llama 3.2 1B. In Figure 9c we show that absorption occurs in these SAEs as well.

Our metric cannot capture absorption past layer 17 in Gemma 2 2B since we rely on ablation experiments to be certain the absorbed feature causally mediates model behavior. Past layer 17, attention has already moved the starting letter information from the source token into the final token position, so any ablations on the source token past layer 17 have little effect. This is a limitation of our absorption metric - we rely on ablation to be certain of the causal impact of absorbed features on model behavior, but this limits the layer depth our metric can be applied. We discuss this further in Appendix A.12 and discuss alternative formulations of the metric in Appendix A.13.

Our absorption metric is not perfect, and is likely an under-estimate of the true level of feature absorption. We only consider absorption to have occurred if a single SAE latent has a much larger ablation effect than all other latents, and if the main SAE latents for a feature do not activate at all. Our metric will not capture multiple absorbing latents activating together, or the main latents activating weakly. Regardless, we feel our metric is a reasonable conservative baseline.

6 Related work

Applications of Probes and SAEs for Model Interpretability Probing methods can extract interpretable information from language models, though this does not guarantee the model uses these representations in its computation, and requires labeled data [7].

Prior work has shown that many human-interpretable concepts in LLM activations are represented as linear directions in activation space, known as the linear representation hypothesis [8, 28]. Li et al. [18] used non-linear probes to recover board representations from a transformer trained on Othello scripts ("OthelloGPT"). Nanda [23] later showed that linear representations were not only recoverable but also editable.

Karvonen et al. [16] developed objective metrics for SAE evaluation using Chess and Othello board states, but does not apply these to SAEs trained on LLM activations. Work by Olah et al. [26], Kissane et al. [17], Templeton et al. [31] noted poor precision/recall of SAE latents compared to known proxies. We extend this by showing how sparsity mediates precision/recall across many Gemma Scope SAEs and offer a possible explanation of low recall due to feature absorption.

Engels et al. [9] investigated SAE errors, finding that not all SAE error is linearly decomposable.

Studying precision and recall of SAE Latents Most existing work on SAE interpretability mainly studies max activating examples [5], which may be misleading. There are more rigorous works which only measure precision [2, 31, 17]. Recent work has briefly explored recall and found it to be worse than expected naively, but this remains poorly understood [26]. We build on this work by evaluating precision / recall on a large number of SAEs, and offer a partial explanation for lower-than-expected recall of SAE latents in the form of feature absorption.

Decomposing SAE Latents Feature splitting was first described in Bricken et al. [2], which noted that different SAE widths and sparsities induce latents of different granularity, with wider SAEs often learning more specific variants of features. Bussmann et al. [4] find that by training an SAE on the decoder of another SAE, a technique called Meta-SAEs, it is possible to break down a single SAE latent like "Einstein" into subcomponents like "German" and "Physicist" and "starts with E".

7 Discussion

Limitations Our Absoption metric uses ablation effect to ensure that the absorbed features causally mediate model behavior, and thus might not be easily transferable to the final model layers. Alternate metric formulations mitigating this are discussed in Appendix A.13. Due to compute constraints, we only train and evaluate a small number of non-JumpReLU SAEs in Figure 9. As our goal was only to show absorption occurs in all SAE architectures, we did not feel this is a significant drawback.

Future Work The primary goal of future work is to find solutions to feature absorption. We are particularly hopeful that work extending Meta-SAEs [4] may solve or mitigate feature absorption. Another possible solution may be attribution dictionary learning [25]. Finally, structured sparsity techniques such as group lasso [13] or hierarchical sparse coding [14] may also be a promising direction of future work.

Other possible directions include allowing absorption to occur and using it as a way to recover hierarchies between features in a LLM. Our toy model results suggest that absorption leads to an asymmetric pattern in the encoder and decoder of the SAE, so it may be possible to use this insight to detect absorption (although there may be other reasons for an asymmetry in the SAE encoder and decoder beyond absorption).

Conclusion We identify a form of feature splitting we call "feature absorption", where more specific latents "steal credit" from more general ones. Absorption creates an interpretability illusion, where a seemingly interpretable latent has arbitrary false negatives in its mainline interpretation. Lower recall poses problems for using SAEs for high-stakes classification or finding sparse circuits [21], as the number of latents needed to characterize model behavior may be much larger than expected.

We show that absorption is a consequence of hierarchical co-occurrence between sparse and dense features. If a dense feature like "starts with letter D" always co-occurs with a more sparse feature like "dogs", the SAE can increase sparsity by absorbing the "starts with D" feature into a "dogs" latent.

We hope that our work highlights the fundamental limitations of sparse feature extraction and prompts future research on SAEs such as identifying cases where a feature "should have activated" but does not due to absorption, and exploring theoretical solutions to absorption. The ease of demonstrating absorption in toy models makes it easier to validate potential solutions.