We use these interventions as part of our absorption metric to ensure that when we claim that "absorption" is occurring, we verify that the absorbing latent has a causal impact on model outputs. This is stronger evidence than only noting a cosine similarity between the absorbing latent, but this means that our absorption metric cannot classify absorption at later model layers.

During a LLM forward pass, the model first collects relevant information on a token in that token position, and attention heads then move relevant information from earlier tokens to later tokens [11, 22]. If we assess ablation effect at layers after which model attention has already pulled relevant information from the subject tokens into the final output token, the ablation effect will be 0. For Gemma 2 2B on the first-letter spelling task, we find this movement of first-letter spelling information occurs around layer 18.

Figure 22 shows an activation patching experiment [22] on a sample first-letter spelling prompt. In this experiment, we see that near layer 18 the model moves first-letter spelling information from the subject token to the prediction token.

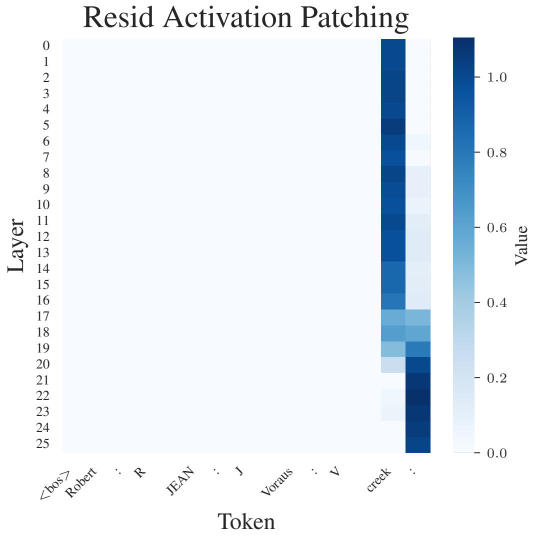

Figure 22: Residual stream attribution patching for a sample first-letter spelling prompt, Gemma 2 2B. After around layer 18, model attention moves the relevant spelling information from the source token to the prediction location.

As a result, our feature absorption metric will not function past layer 18 in Gemma 2 2B, and we thus focus on layers 0-17 for our analysis of feature absorption. We believe that feature absorption is still occurring in SAEs past layer 18, but we lose the ability to make causal claims that the absorbing latents are used by the model to make predictions. Given that this paper is trying to highlight the existence of feature absorption, we felt it is more important to have a metric which is robust and has the backing of causal analysis but which cannot be used at all model layers. Future work may make a different trade-off and choose a feature absorption metric which can work at all model layers, for instance relying only on cosine similarity between absorbing latents and a LR probe to determine absorption. We describe variations on the metric that can facilitate this in Appendix A.13 below.

A.13 Alternate formulations of the absorption metric

There are a few variations to the metric that can be made to make it more flexible so that it can be applied at all layers of the model, and can be tweaked to detect partial absorption as well. As described above, we did not use this variation as we felt it is more important to demonstrate definitely and conservatively that absorption occurs in this paper. However, we describe the changes that can be made to the metric as follows:

replace ablation study with LR probe projection in reconstruction The metric in our paper uses an integrated-gradients ablation study to be absolutely certain that an absorbing feature is causally responsible for model behavior. We use this in the paper as our goal is to establish with certainty that feature absorption exists in real models, but it has the following drawbacks:

- We need to be able to consistently prompt the model to perform the task, outputting a single token
- We cannot evaluate final model layers after attention moves task-relevant information to the final token
- Ablation studies are slow, which makes the absorption metric expensive to calculate.

The ablation study can be replaced instead with a threshold on the portion of the logistic regression (LR) probe direction an absorbing latent contributes to the residual stream. To do this, we project all firing latents against the LR probe direction d_p, as well as project the input activation a against the probe. We require that a latent l must contribute at least τ_c portion of the probe projection to the reconstruction. So, in order for l to be considered an absorbing latent, the following must be true:

$$\tau_c < \frac{\hat{a}_l \cdot d_p}{a \cdot d_p}$$

where $\hat{a}_l$ is the reconstruction component of latent l (the encoder activation of latent l times the decoder vector for l).

Allow multiple absorbing latents The metric in the paper requires one dominant absorbing latent for simplicity, but it is possible in theory to have two or more absorbing latents firing together performing absorption. We can modify the metric to take this into account by allowing up to N latents firing together contribute up to τ_c together.

Thus, the threshold from change 1 above now becomes the following, where l_n refers to the absorbing latent with the nth largest contribution to the LR probe direction in the reconstruction:

$$\tau_c < \sum_{n}^{N} \frac{\hat{a}_{l_n} \cdot d_p}{a \cdot d_p}$$

We do not do this in the paper as it requires another hyper-parameter which is hard to set in a principled way, but should result in a less conservative estimate of absorption.

Allow main latent(s) to fire weakly instead of being fully turned off The metric in the paper requires that the main latent(s) for a task all be fully disabled to identify a case of absorption. However, in partial absorption, as explored in our section on toy models, the main latent does not always fully turn off but fires very weakly instead. We can adjust the metric to take this into account by relaxing the requirement that the main latents are fully disabled and instead allow them to fire weakly. Similar to change 1, we define a threshold τ_m as a maximum contribution to the probe direction in the reconstructed activation $\hat{a}$ that comes from each main latent l_m. Thus, if we have M main latents, in order to be classified as absorption the following must be satisfied:

$$\tau_m \geq \sum_{m}^{M} \frac{\hat{a}_{l_m} \cdot d_p}{a \cdot d_p}$$

In the metric defined in the paper, $\tau_m = 0$, meaning the main latents must be fully turned off to count as absorption.

This final change, which detects partial absorption, may be preferable to get a sense of an overall level of absorption present in an SAE. However, if the goal is to determine whether absorption affects the SAE's ability to act as a classifier, then requiring that the main SAE latent be fully turned off to be called absorption is preferable, so we do not claim that either version of the metric is superior in all cases.

A.14 Additional plots

In this section, we include additional plots that are too large to fit in the main body of the paper.

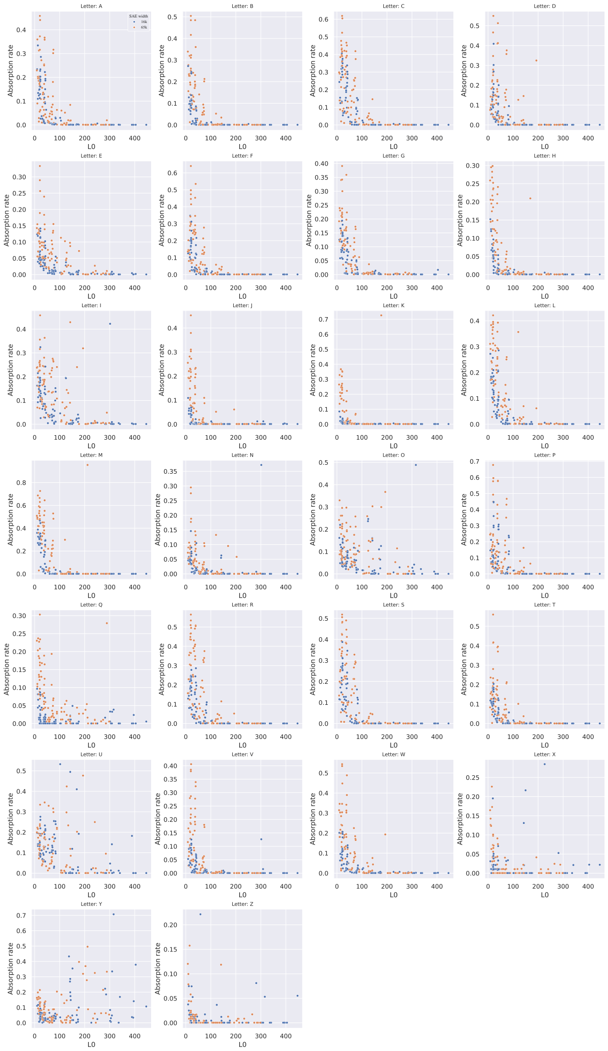

Figure 23: Absorption rate vs L0 by letter, layers 0-17. We see a wide variance in which letters are absorbed by which SAEs.