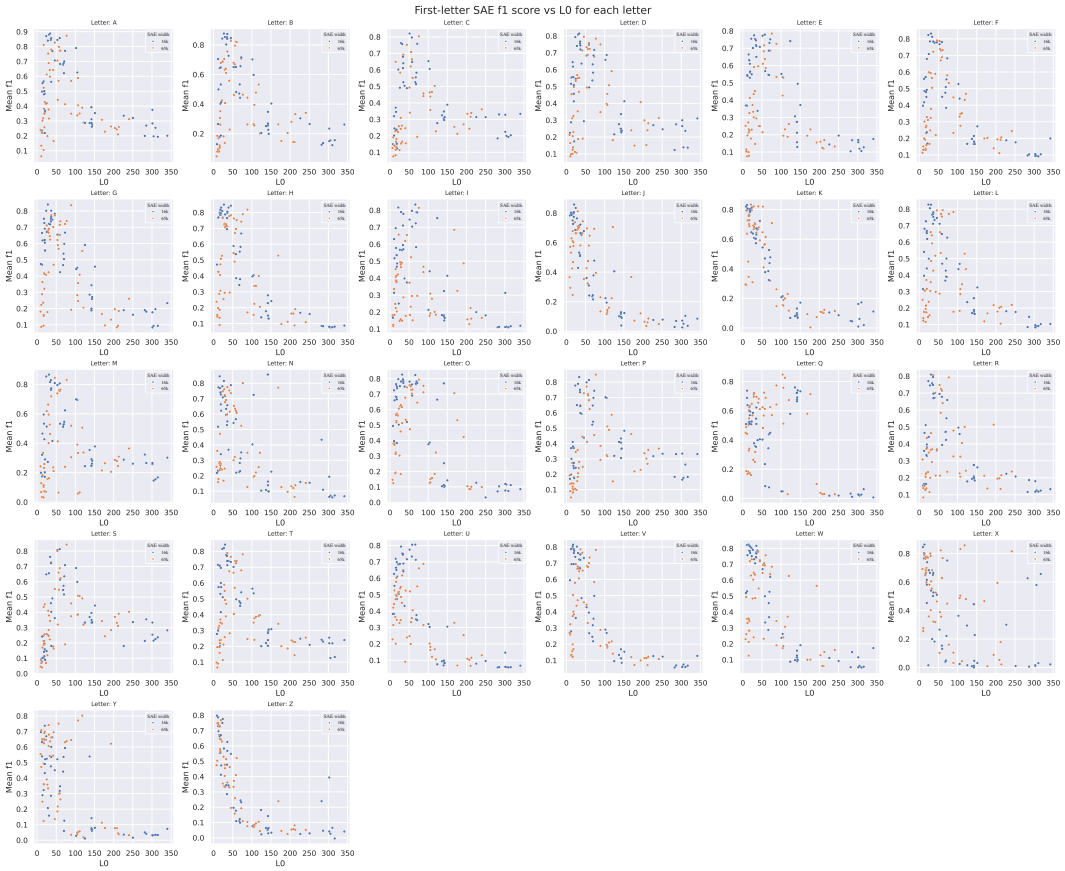
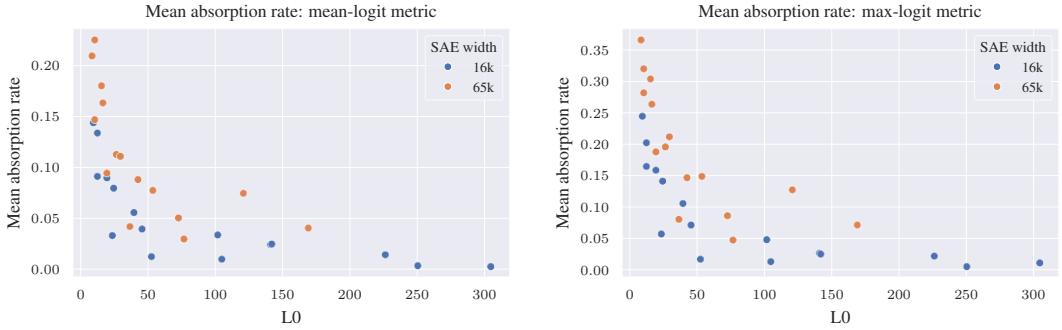


Figure 20: F1 vs L0 by letter. SAE latents are picked using k=1 sparse probing.

$$m_{max} = g[y] - \max_{l \in \{L \setminus y\}} g[l]$$

This second form using a max can also account for the case where the logits of the model shift from being confident in the correct answer to instead being confident in an incorrect answer while leaving the logits of the correct answer the same.



(a) Absorption rate using the mean version of the absorption metric, Gemma Scope layers 0-3.

(b) Absorption rate using the max version of the absorption metric, Gemma Scope layers 0-3.

Figure 21: Comparison of absorption rates using the max and mean versions of the absorption metric.

In practice, we expect that ablating an absorbing latent should cause the model to become less confident in the correct answer, so the difference between these two forms of the metric should yield similar results.

We calculate the mean absorption rate for Gemma Scope SAEs layers 0-3 in Figure 21 using both versions of this metric. The overall shape of the curve is nearly identical between these two choices of metrics. The mean version of the metric, which is used in this paper, results in a slightly more conservative estimate of absorption rate.

We consider our absorption score to be a rough estimate of the true absorption rate and thus consider either the mean or the max version of the logit diff metric to be valid for evaluating absorption.

A.11 Choice of thresholds for absorption metric

Our absorption metric makes use of several thresholds. For feature splitting, we use a threshold of 0.03 on the F1 score jump moving from $k = n$ to $k = n + 1$ in sparse probing. To classify a latent as a case of absorption, we require cosine similarity with the LR probe above 0.025, and a clear largest ablation effect of 1.0 more than then next highest latent. In this section, we discuss the intuition behind these thresholds. In all cases, these thresholds are just the rough midpoint of ranges of values that achieve similar results, and that we feel are reasonable default values. We also refer readers to our online feature absorption explorer to gain a similar intuition as to what typical ranges of these values look like.

Feature splitting threshold We detect interpretable feature splitting by noting how large a jump in F1-score we gain moving from $k = n$ to $k = n + 1$ in sparse probing. Figure 8a demonstrates a typical example of how F1 score increases for interpretable feature splitting vs feature absorption. In the left side of the figure, we see that moving from $k = 1$ to $k = 2$ there is a F1-score jump of around 0.08, while each increase after that is less than 0.01. For the figure on the right, where no interpretable feature splitting occurs, all F1-score jumps are less than 0.01. This plot is a very typical illustration of detecting feature splitting via k-sparse probing. Any threshold between 0.01 and 0.05 does a good job of detecting feature splitting. We set the threshold to 0.03 to be in the middle of this range.

LR probe cosine similarity threshold We use a threshold of 0.025 on the cosine similarity of the firing latent with the LR probe as part of the metric, ensuring that any latent we classify as absorption must contain some component of the probe direction. This threshold is mainly a cheap way to filter out latents that are obviously not probe-aligned so we can avoid running the more expensive ablation experiments. Nearly any threshold above 0 and below 0.05 should work identically well for this purpose. We choose 0.025 to be in the middle of this range. For most cases of absorption we detect, the absorbing latent has a probe cosine similarity of around 0.05 - 0.15. Figure 6b demonstrates a very typical case of cosine similarity between an absorbing latent and the LR probe, showing cosine similarity of 0.12.

Absorption effect threshold As part of our metric, we assert that any latent we classify as absorption must have the highest ablation effect of all latents, and that lead must be by at least 1.0. As the main goal of the metric in our paper is establishing definitively that absorption exists and affects real LLM SAEs, we focus on the most obvious cases of absorption where a single latent absorbs the parent feature direction in a single activation. This threshold thus serves two purposes: first, to establish the latent has a strong ablation effect, and second, the establish that the effect is dominant over other latents. The metric is not particularly sensitive to the exact choice of threshold we set here, as any dominant ablation effect above around 0.5 is already quite strong. Figure 7 shows a typical ablation effect for a dominant absorbing latent, and has an ablation effect above 6, while all other latents have ablation effects well below 0.5. We pick a threshold of 1.0 here as its a clean number that is well within the range that will work as a threshold. Any threshold between 0.5 and 3.0 should work just as well.

A.12 Causal interventions and absorption

In this work, we rely on causal interventions like ablation experiments to verify that SAE latents have a causal impact on model behavior. In these experiments for spelling tasks, we set up an ICL prompt to elicit spelling information from the model, for instance the ICL prompt below:

```
tartan has the first letter: T
mirth has the first letter: M
dog has the first letter:
```

In this ICL prompt, we would apply an SAE and train LR probes on the `_dog` token position, and expect that the model will output the token `_D`. When we intervene on the `_dog`, we can track the causal changes to model outputs by applying a metric to the output logits, e.g. checking how our intervention increases or decreases the `_D` logit relative to other letters.