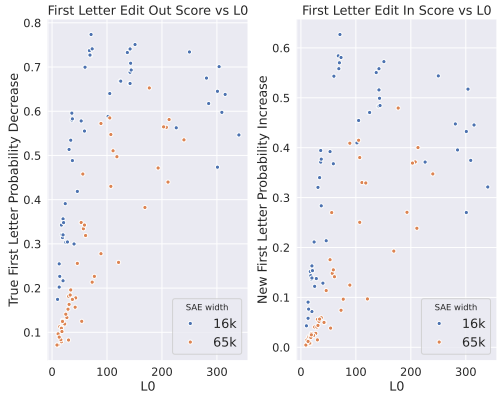
(a) Comparing success in editing out the true first letter and making the model predict a randomly selected new letter across layers 0-9 for all 16k and 65k Gemma Scope SAEs.



(b) Comparing the edit success with the top SAE latent across all L0s for 16k and 65k widths across layers 0-9. The best performance seems to be occurring for L0 between 75 and 150.

Figure 14: Comparison of Edit success by Layer and L0

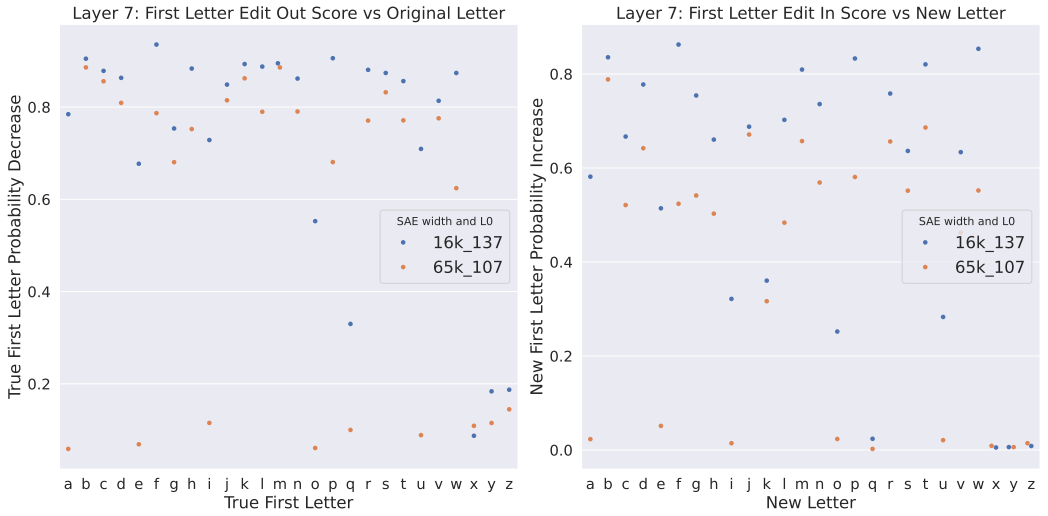


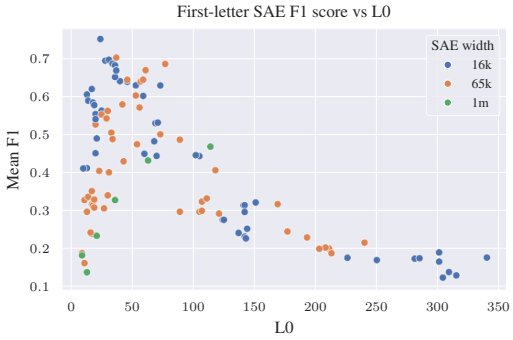
Figure 15: Comparing the edit success broken down by the letter at layer 7 for two SAEs; SAE width 16,000 and L0 of 137 and SAE width 65,000 and L0 of 107. For each original letter we draw a sample of 100 tokens and average the decrease in probability of the correct first letter and increase in probability of a new random letter.

latent. Likely any reasonable method of latent selection will find the same latent for these cases. We thus decided to use cosine similarity between the SAE encoder and a LR probe as our selection criteria for single SAE latents as this is a simpler metric and less computationally intensive to compute.

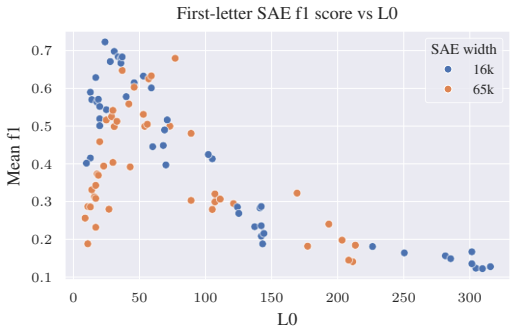
A.8 Precision, recall, and F1 score for the first-letter task

We evaluated precision, recall, and F1 score for the first-letter classification task, and found that the precision and recall vary depending on the L0 of the SAE. Low L0 SAEs learn high precision, low recall latents, while high L0 SAEs learn low precision, high recall latents. These results are shown in Figure 19. We thus chose to use F1 score as our core metric in this paper to balance precision and recall as many if the SAEs we tested have extreme values in either precision or recall.

While it may appear that there is an optimal L0 from looking at aggregate statistics across letter, we find that breaking down the F1 vs L0 plot by letter reveals that the optimal L0 appears different for different letters, with low frequency letters like z actually having the best F1 score at the lowest L0,



(a) Mean F1 score on first-letter classification task using top SAE encoder latent by cosine similarity with the LR probe vs L0 for all Gemma Scope SAEs layers 0-9.



(b) Mean F1 score on first-letter classification task using k=1 sparse probing to select the SAE latent for the first-letter classification task vs L0 for all Gemma Scope SAEs layers 0-9.

Figure 16: Comparison of LR probe cosine similarity and k=1 sparse probing vs l0

(a) Mean F1 score on first-letter classification task using top SAE encoder latent by cosine similarity with the LR probe vs layer for all Gemma Scope SAEs layers 0-9.

(b) Mean F1 score on first-letter classification task using k=1 sparse probing to select the SAE latent for the first-letter classification task vs layer for all Gemma Scope SAEs layers 0-9.

Figure 17: Comparison of LR probe cosine similarity and k=1 sparse probing vs layer

while other letters instead have an optimal L0 around 30-50. Figure 20 shows these results broken down by letter.

A.9 SAE training

We train SAEs on the first 8 layers of Qwen2 0.5B [32] and Llama 3.2 1B [6] using the SAELens library [15]. The SAEs are all trained with identical hyperparameters of L1 coefficient of 2.5 and 500M tokens. The Qwen2 0.5B SAEs all have L0 between 25 and 50 and explained variance between 0.77 and 0.83. The Llama 3.2 1B SAEs have L0 between 27 and 110, and explained variance between 0.74 and 0.89. We use a single 40gb Nvidia A100 GPU for training each SAE.

A.10 Metric choice for ablation studies

To determine the causal effect of SAE latents on the first-letter identification task, we use a metric, m, which measures the logit of the correct letter minus the mean logit of all incorrect letters. Our metric is defined below, where g refers to the final token logits, L is the set of uppercase letters, and y is the uppercase letter that is the correct starting letter:

$$m = g[y] - \frac{1}{|L| - 1} \sum_{l \in \{L \setminus y\}} g[l]$$

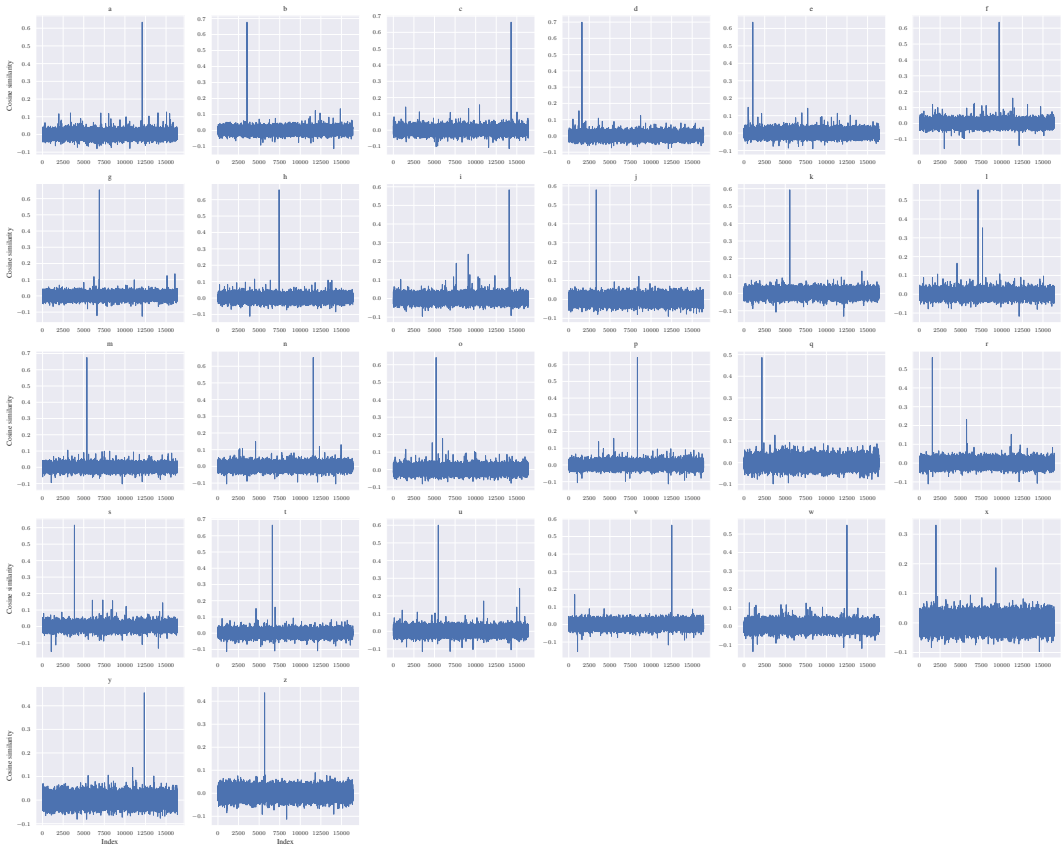
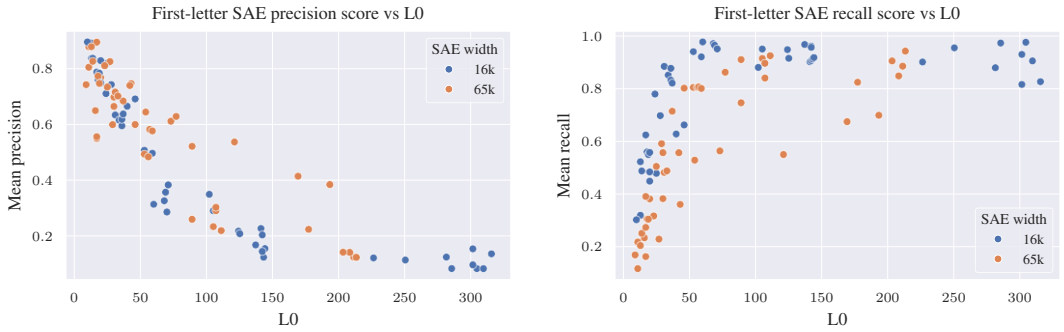


Figure 18: Decoder cosine similarities with the LR probe by letter, Gemma Scope 16k layer 0 l0=105. Most letters have one or two obvious SAE latents which align with the probe.



(a) Mean precision on first-letter classification task vs L0 for all Gemma Scope SAEs layers 0-9. Latents are selected via k=1 sparse probing

(b) Mean recall on first-letter classification task vs L0 for all Gemma Scope SAEs layers 0-9. Latents are selected via k=1 sparse probing

Figure 19: Comparison of precision and recall vs l0

This metric is chosen to detect changes in the confidence of the model in predicting the correct letter relative to the mean reference class of other letters. This should capture changes in the model's confidence in predicting the correct logit.

This is not the only metric that could be chosen, and an argument can be made that we should subtract the max of all incorrect letter logits rather than the mean of all incorrect letter logits. The max form of this version of the metric is shown below: