

Figure 13: ROC curves for ‘detection’ auto-interpretability for Pythia-1b over 100 SAE latents. These results demonstrate the similarity in performance between the SAE variants, although here we do not observe an overall degradation in quality.

## B.5 PYTHIA 6.9B

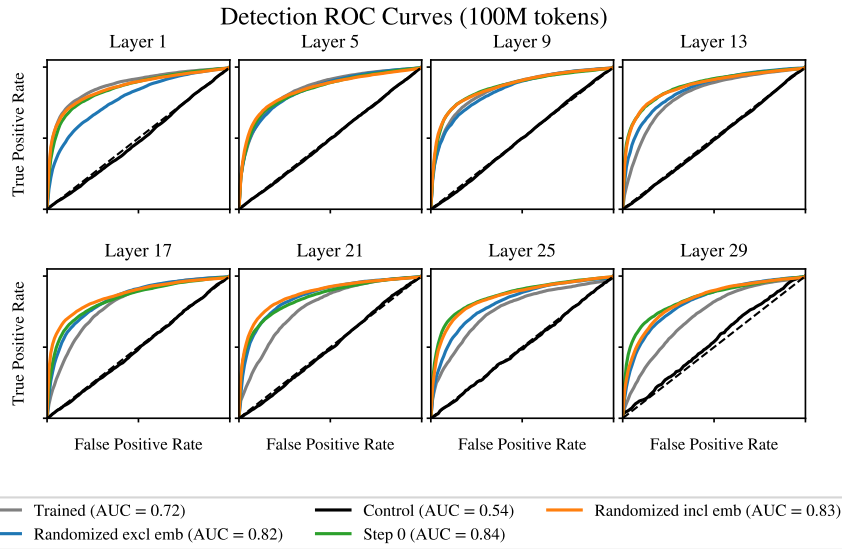


Figure 14: ROC curves for ‘detection’ auto-interpretability for Pythia-6.9b over 100 SAE latents. These results demonstrate the similarity in performance between the SAE variants.

## C EFFECT OF INCREASED TRAINING DATA

For our primary experiments, we trained SAEs on 100M tokens (Section 3). We verified that our results were not explained by a lack of sufficient training data by repeating a subset of these experiments with SAEs trained on 1B tokens from the RedPajama dataset (Figure 15).

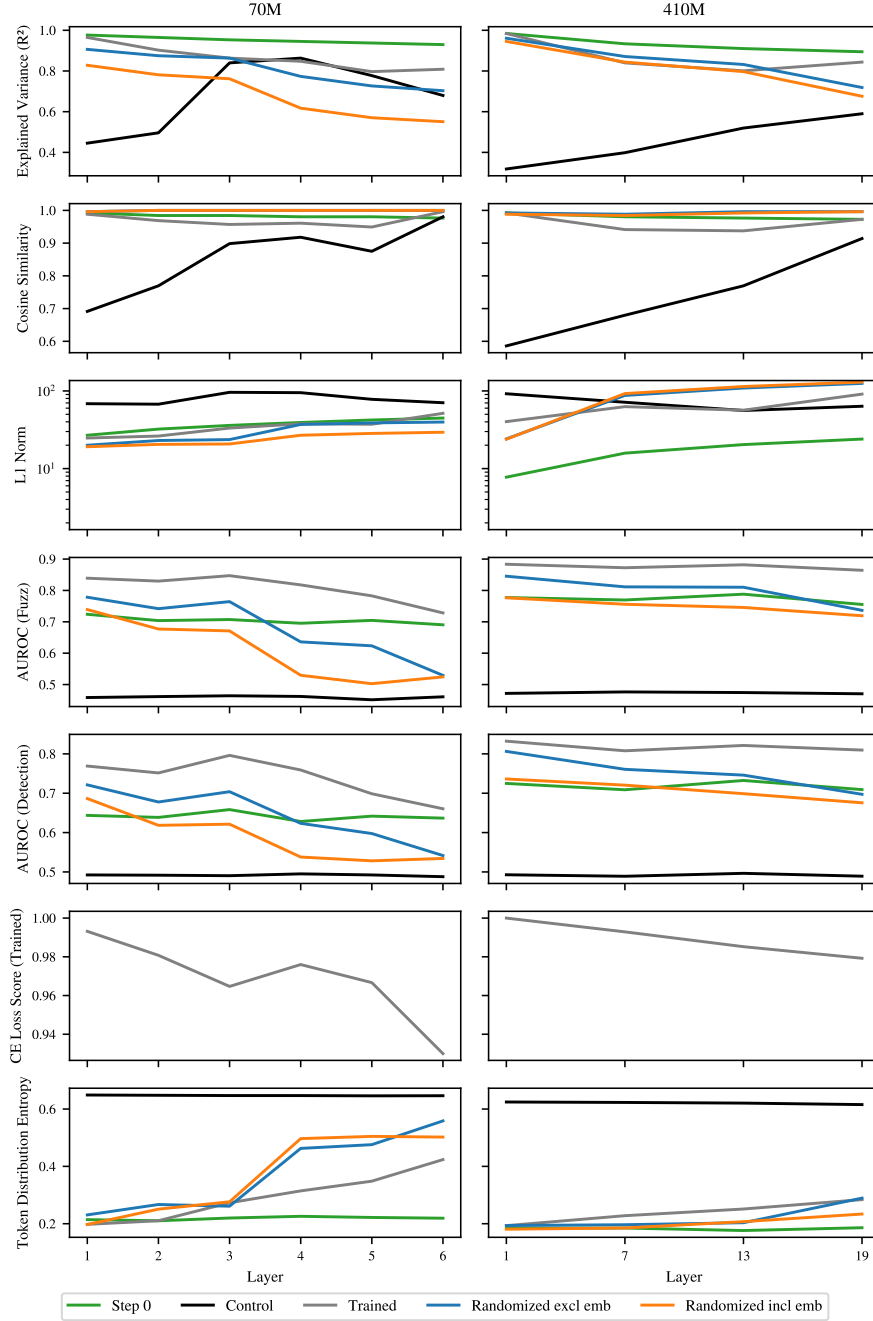


Figure 15: Evaluation metrics for SAEs trained with one billion tokens on the Pythia-70m and 410m models. These results correspond to columns of Figure 2, which show the same evaluation metrics for SAEs trained on 100M tokens, and qualitatively similar behavior.

## D EFFECT OF DECREASED TRAINING DATA FOR PYTHIA-1B

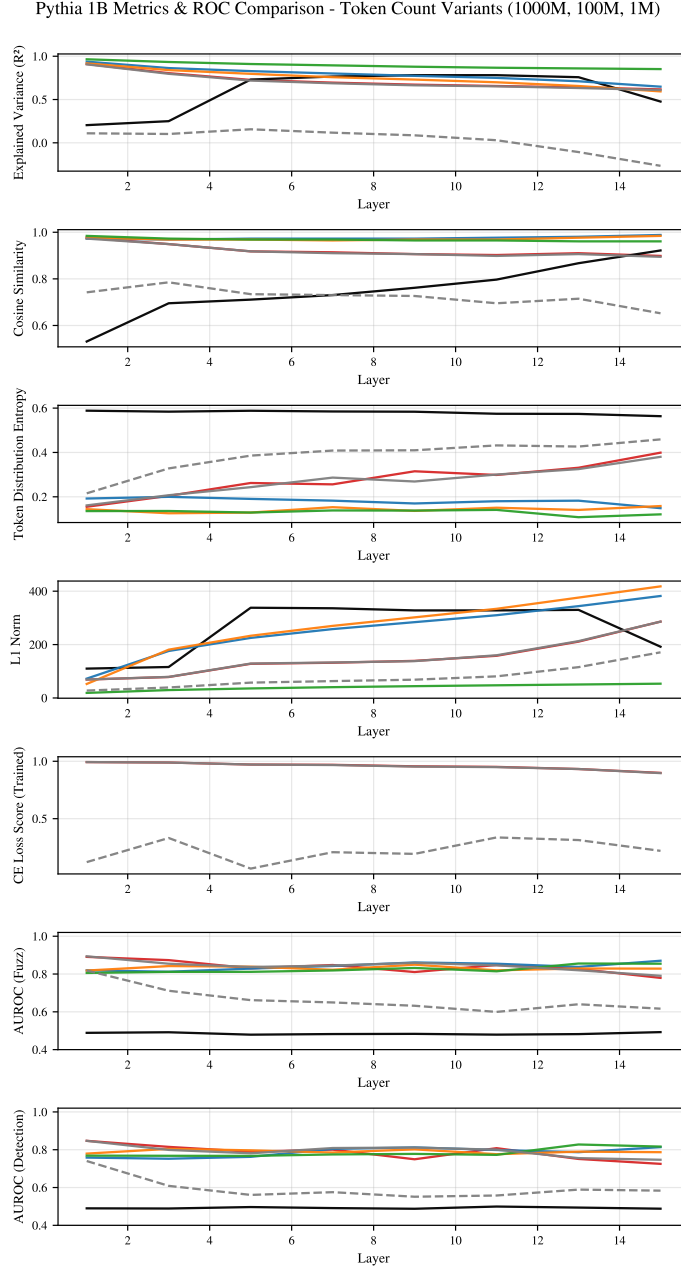
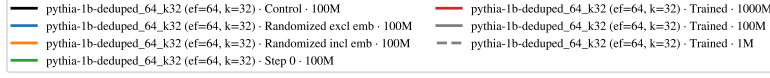


Figure 16: Evaluation metrics for SAEs trained with 1M and 1B tokens on Pythia-1b. The explained variance and CE loss score are significantly lower for the 1M model, showing that the SAEs are under-trained. Average auto-interpretability scores are slightly lower for the earliest layers, but decline sharply with increasing layer. The trends in auto-interpretability and token distribution entropy with layer index are consistent with other SAEs.