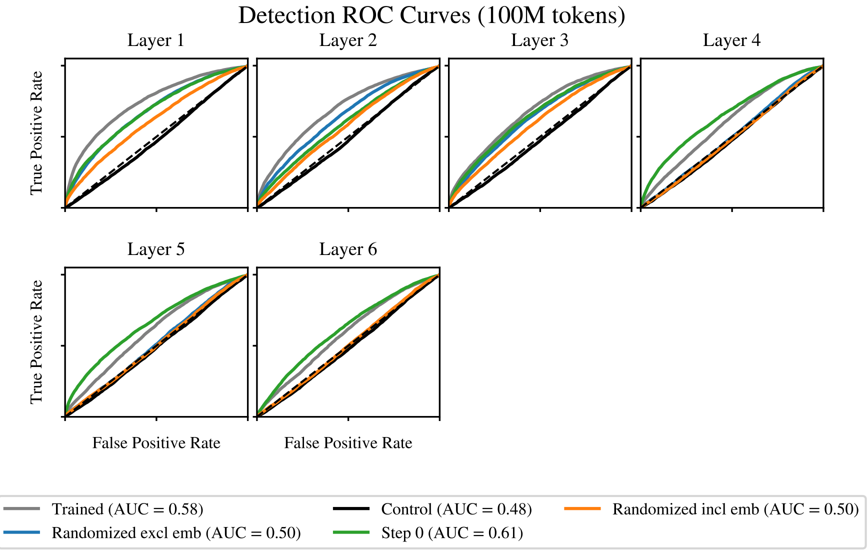

Figure 7: ROC curves for ‘detection’ auto-interpretability for Pythia-70m over 100 SAE latents. These results demonstrate the similarity in performance between the SAE variants, as well as the overall degradation in performance as the layer index increases.

### B.2 Pythia 160m

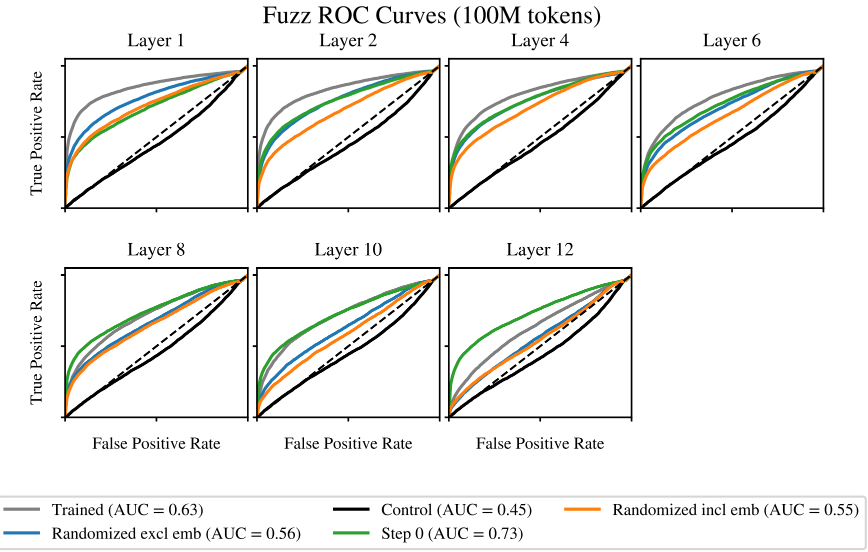

Figure 8: ROC curves for ‘fuzzing’ auto-interpretability for Pythia-160m over 100 SAE latents. These results demonstrate the similarity in performance between the SAE variants, as well as the overall degradation in performance as the layer index increases.

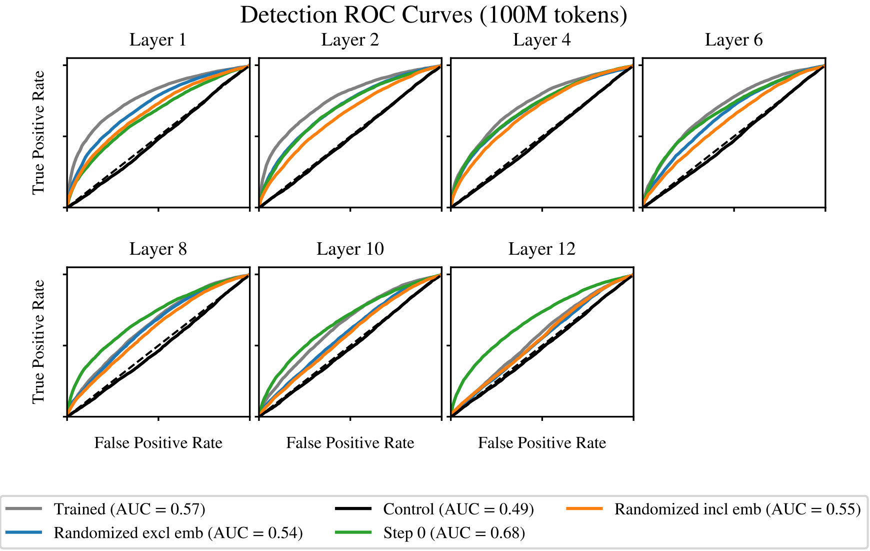

Figure 9: ROC curves for 'detection' auto-interpretability for Pythia-160m over 100 SAE latents. These results demonstrate the similarity in performance between the SAE variants, as well as the overall degradation in performance as the layer index increases.

### B.3 Pythia 410m

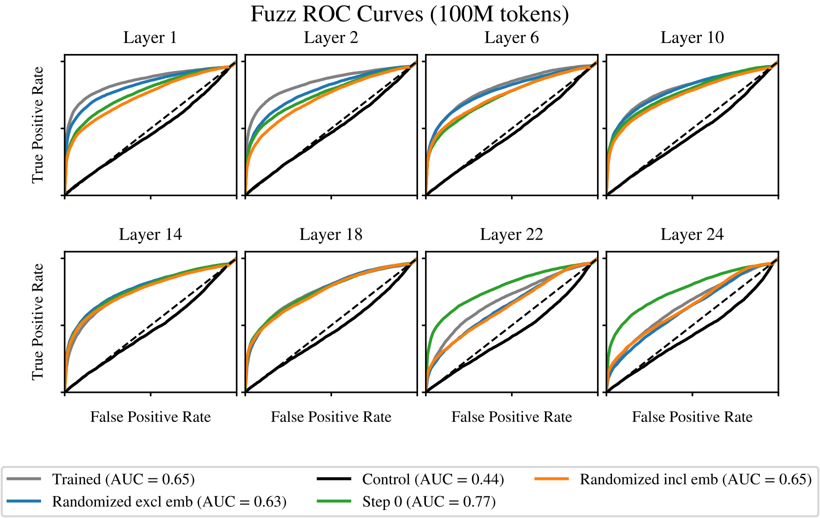

Figure 10: ROC curves for 'fuzzing' auto-interpretability for Pythia-410m over 100 SAE latents. These results demonstrate the similarity in performance between the SAE variants, as well as the overall degradation in performance as the layer index increases. The auto-interpretability scores here fail to distinguish between trained and randomized models.

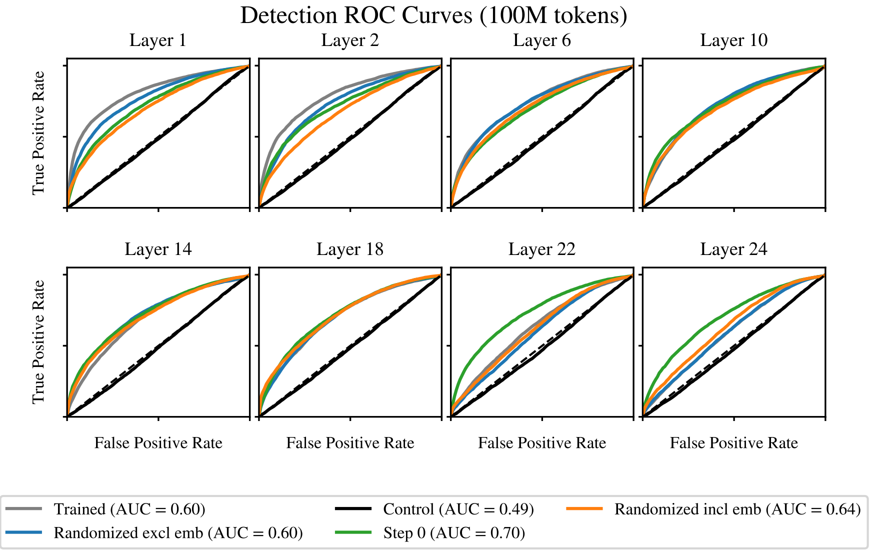

Figure 11: ROC curves for 'detection' auto-interpretability for Pythia-410m over 100 SAE latents. These results demonstrate the similarity in performance between the SAE variants, as well as the overall degradation in performance as the layer index increases.

### B.4 Pythia-1b

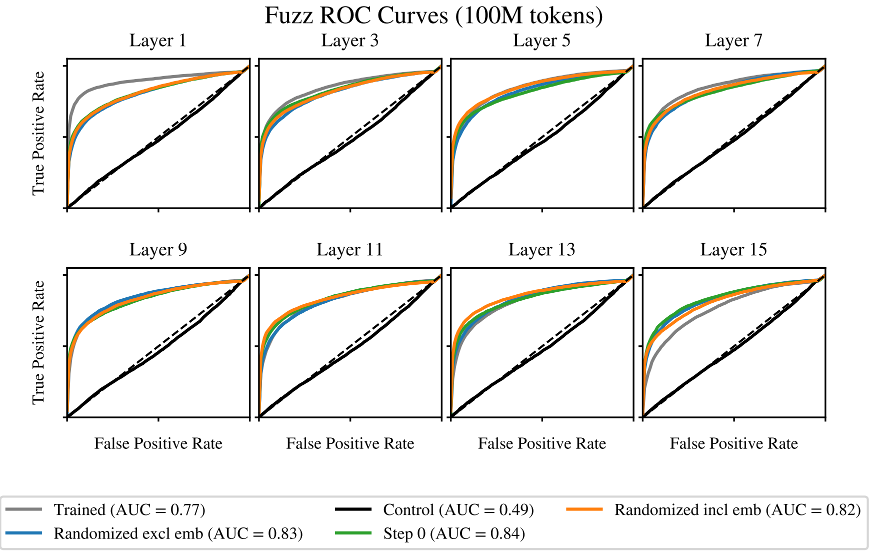

Figure 12: ROC curves for 'fuzzing' auto-interpretability for Pythia-1b over 100 SAE latents. These results demonstrate the similarity in performance between the SAE variants, although here we do not observe an overall degradation in quality.