E UNCERTAINTY PLOTS FOR PYTHIA-70M

We computed uncertainty for our evaluation metrics on Pythia-70m using five random seeds.

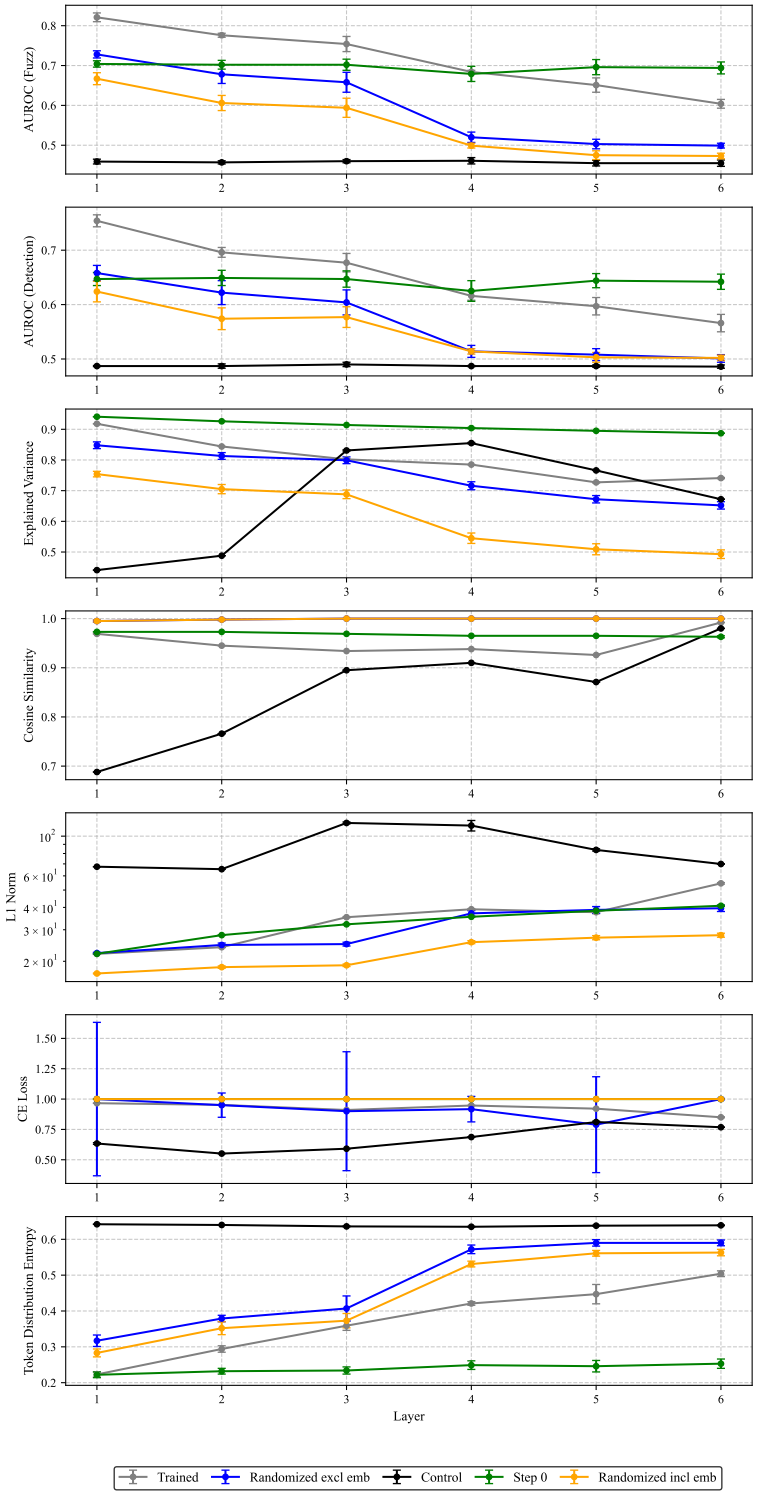

Figure 17: Uncertainty for Pythia-70m metrics computed using five random seeds.

F Effect of SAE hyperparameters for Pythia-160M

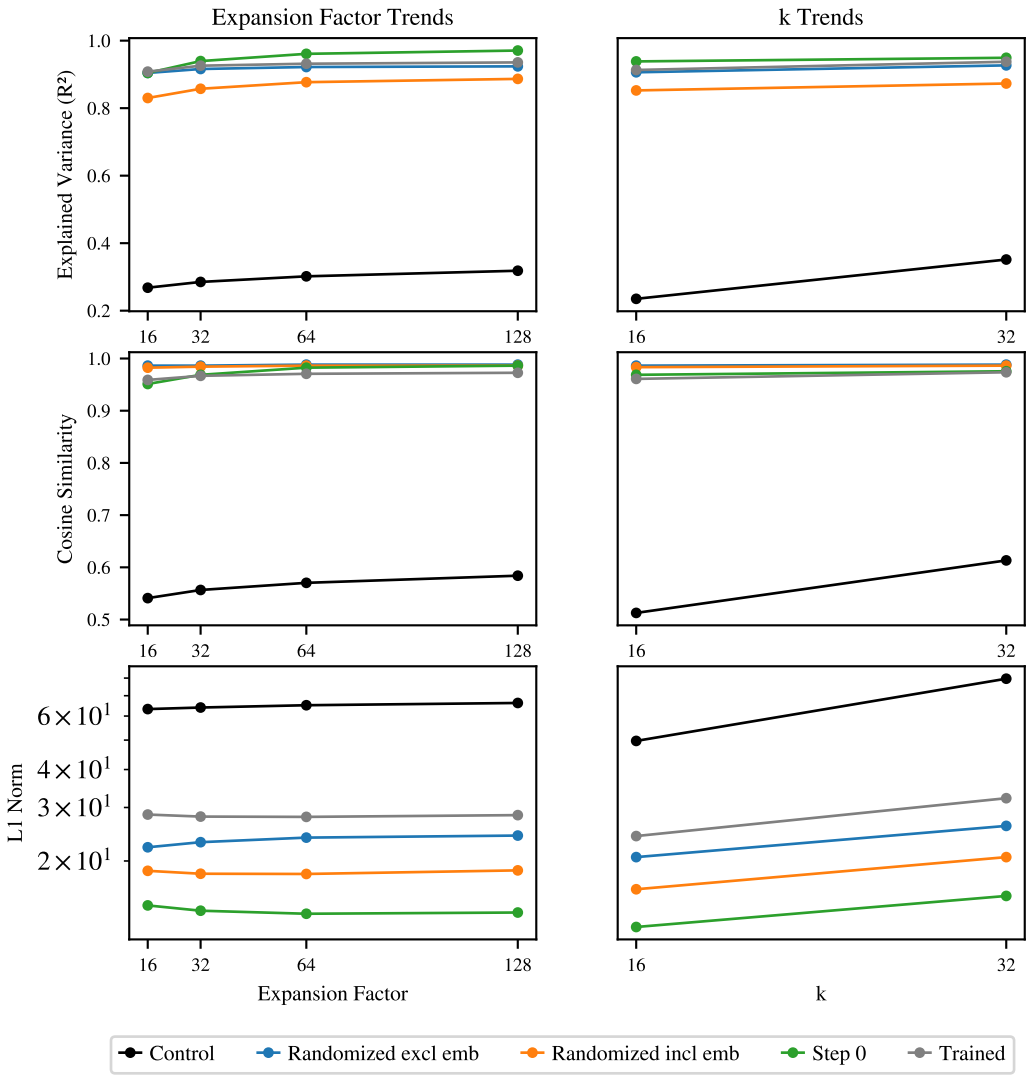

Figure 18: Robustness of SAE performance to hyperparameter selection. Standard evaluation metrics remain stable across a wide range of expansion factors R (16 to 128) and sparsities k (16 to 32), with all initialization strategies maintaining their relative performance ordering. This stability suggests that moderate hyperparameter values (e.g., expansion factor $R = 64$, sparsity $k = 32$) suffice.

G EFFECT OF SAE HYPERPARAMETERS FOR PYTHIA-1B

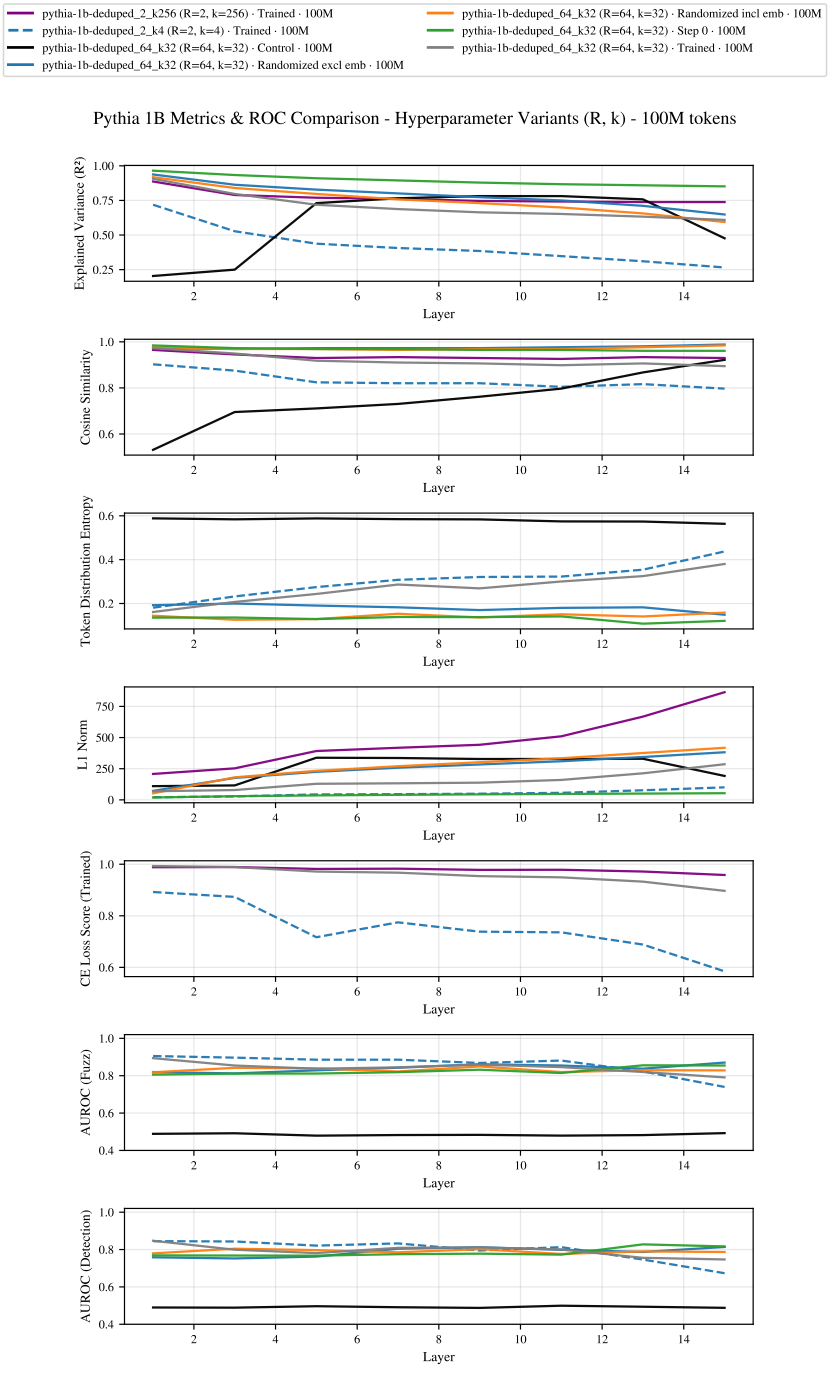

Figure 19: Evaluation metrics for SAEs trained on the Pythia-1b model with different hyperparameters, including the main results from Figure 2. SAEs with a very small expansion factor $R = 2$ and sparsity $k = 4$ are clearly distinguished from our default hyperparameters by the explained variance and CE loss score. Importantly, the auto-interpretability scores of these SAEs remain similar to those trained with default hyperparameters on either trained or randomised models.