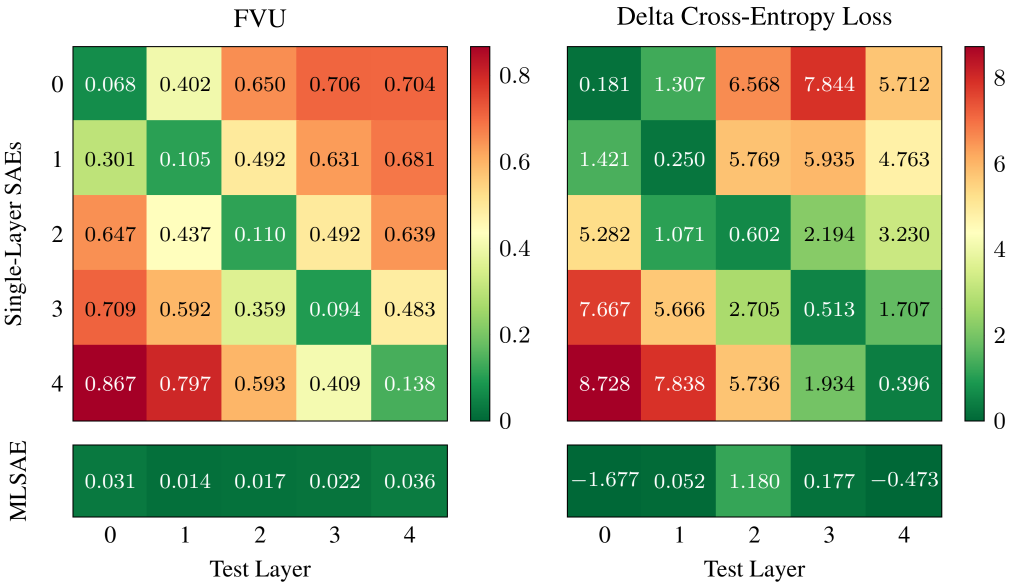

Figure 22: The FVU reconstruction error and delta cross-entropy loss for single-layer SAEs trained on each layer of Pythia-70m compared with a single multi-layer SAE trained on every layer, applying tuned-lens transformations during training and evaluation (Section 3.3).

B.5 Mean Max Cosine Similarity

Sharkey et al. (2022) define the Mean Max Cosine Similarity (MMCS) between a learned dictionary X and a ground-truth dictionary X'. There is no ground-truth dictionary for language models, so a larger learned dictionary or the k nearest neighbors to each dictionary element are commonly used.

$$\text{MMCS}(X, X') = \frac{1}{|X|} \sum_{\mathbf{x} \in X} \max_{\mathbf{x}' \in X'} \cos \text{sim}(\mathbf{x}, \mathbf{x}') \tag{13}$$

The MMCS serves as a proxy measure for ‘feature splitting’ (Bricken et al., 2023; Braun et al., 2024): as the number of features increases, we expect the decoder weight vectors to be more similar to their nearest neighbors. We compute the MMCS with $k = 1$ after training, finding it decreases slightly as the model size increases with fixed hyperparameters (Table 2).

B.6 Pairwise cosine similarities

A potential issue when training multi-layer SAEs is learning multiple versions of ‘the same’ latent, i.e., multiple latents with similar interpretable functions but which are active at different layers. In this case, we would expect to find pairs of latents with relatively large cosine similarities between their decoder weight vectors but different observed distributions of activations over layers (Section 4.3). We investigated this possibility by comparing the distributions of pairwise cosine similarities between decoder weight vectors for trained MLSAEs to reference distributions.

As a negative control, we generated an equal number (the number of latents n) of normal independently and identically distributed (i.i.d.) vectors $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ of the same length (the model dimension d). In this case, the pairwise cosine similarities follow a normal distribution $\cos \text{sim}(\mathbf{x}, \mathbf{x}') \sim \mathcal{N}(0, 1/d)$. As a positive control, we generated a smaller number of normal i.i.d. vectors (the number of latents n divided by the number of layers n_L), copied the vectors n_L times, and added noise $\sim \mathcal{N}(0, 1)$ to each copy. In this case, we expect an additional frequency peak for large positive cosine similarities.

Figure 24 shows that the distributions of pairwise cosine similarities for decoder weight vectors are slightly heavier-tailed and right-shifted compared with the negative control, i.e., a pair of MLSAE latents are slightly more likely to have high cosine similarity than a pair of i.i.d. normal vectors. However, the number of pairs with large positive cosine similarities is small compared to the positive control, which has a second peak around 0.5 (visible only with the logarithmic y-axis scale).

Model	Mean	Std. Dev.
Pythia-70m	0.275	0.0843
Pythia-160m	0.250	0.0928
Pythia-410m	0.221	0.0868
Pythia-1b	0.201	0.0989
Pythia-1.4b	0.180	0.0861
Gemma 2 2B	0.249	0.1052
Llama 3.2 3B	0.215	0.1187
GPT-2 small	0.258	0.0703

(a) Standard

Model	Mean	Std. Dev.
Pythia-70m	0.261	0.0763
Pythia-160m	0.206	0.0734
Pythia-410m	0.216	0.0864

(b) Tuned lens

Table 2: The mean and standard deviation of the maximum cosine similarity between decoder weight vectors for MLSAEs with an expansion factor of $R = 64$ and sparsity $k = 32$. The MMCS decreases as the model size increases for Pythia transformers.

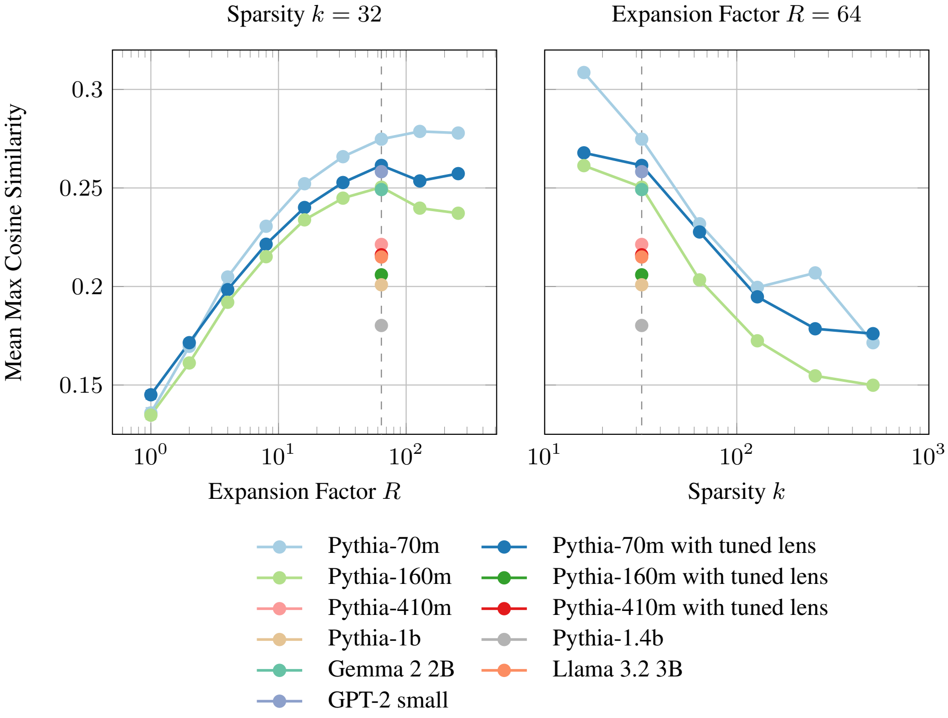

Figure 23: The Mean Max Cosine Similarity between decoder weight vectors for standard and tuned-lens MLSAEs. The MMCS increases as the expansion factor R increases and decreases as the sparsity k increases. Applying tuned-lens transformations slightly decreases the MMCS relative to standard MLSAEs.

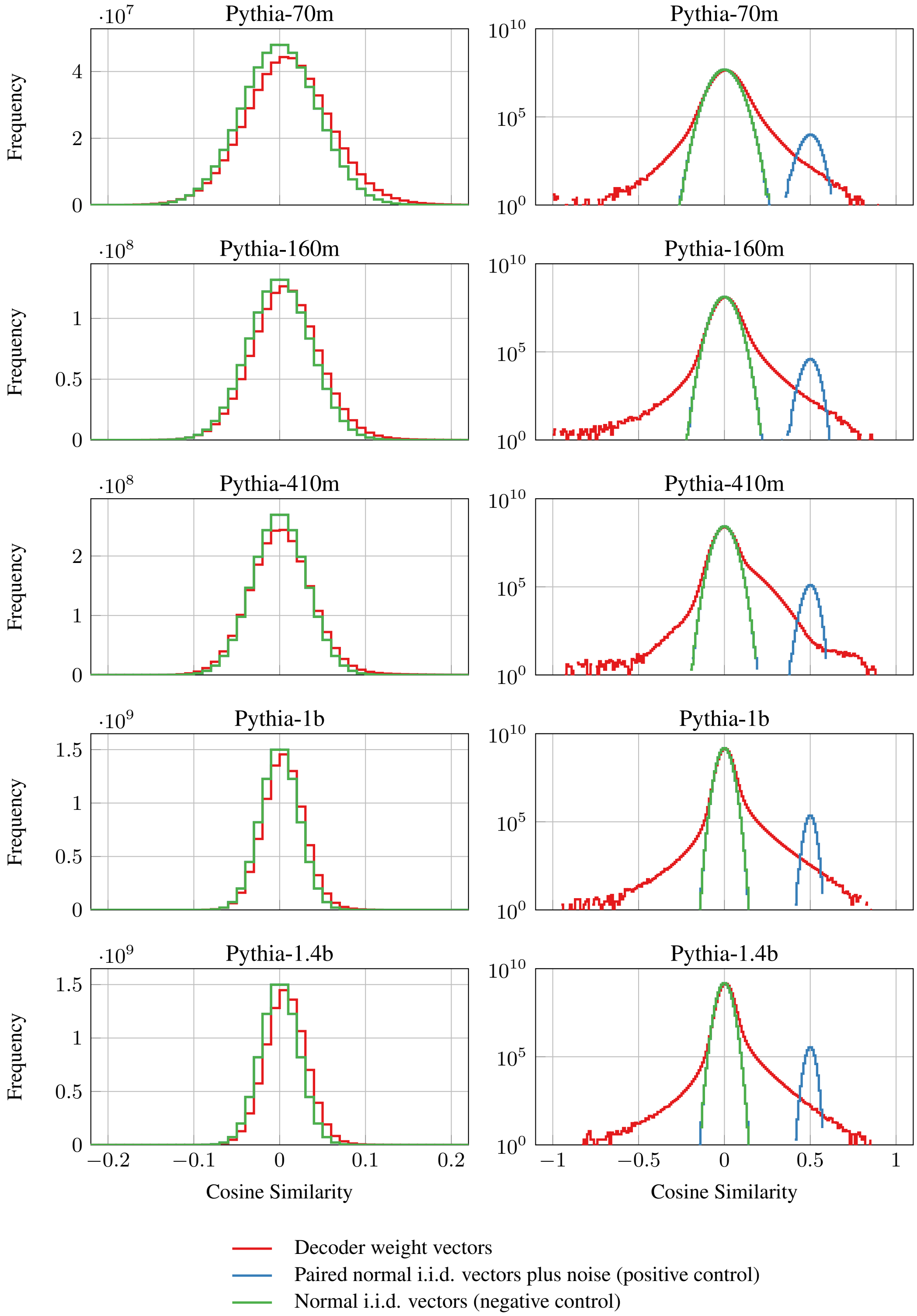

Figure 24: Histograms of the frequencies of pairwise cosine similarities between decoder weight vectors, compared to an equal number of normal i.i.d. vectors of the same length, and n_L copies of a smaller number of normal i.i.d. vectors with added noise $\sim \mathcal{N}(0, 1)$. Here, we report the frequencies for MLSAEs trained on Pythia models with an expansion factor of $R = 64$ and sparsity $k = 32$. The left-hand y-axis scale is linear, the right-hand is logarithmic.