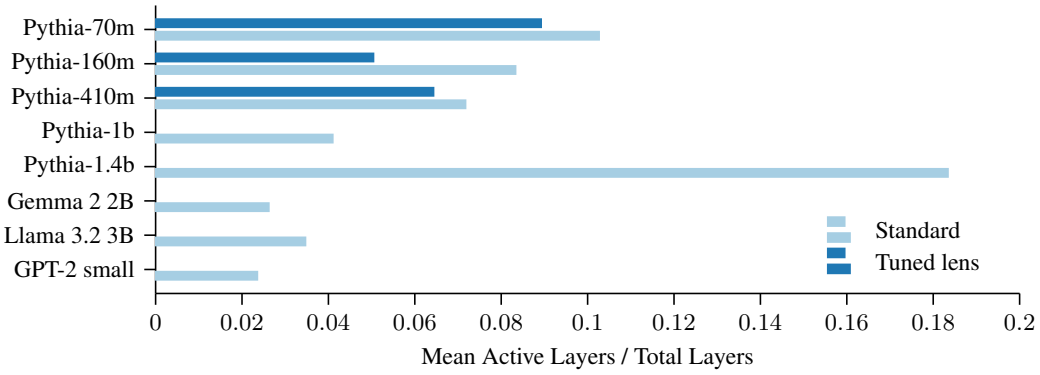
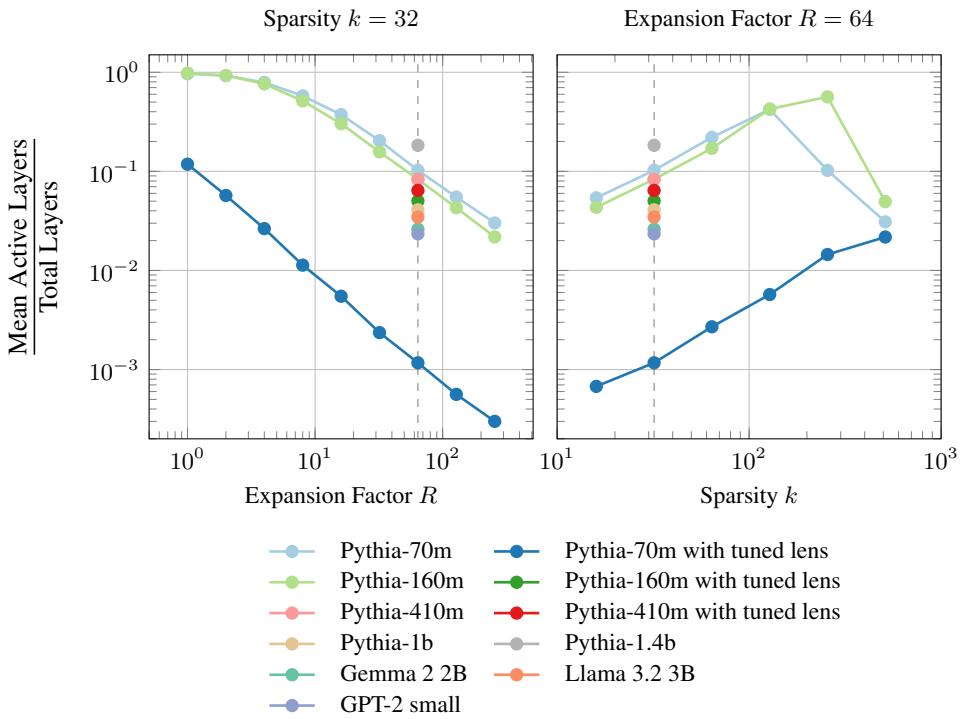
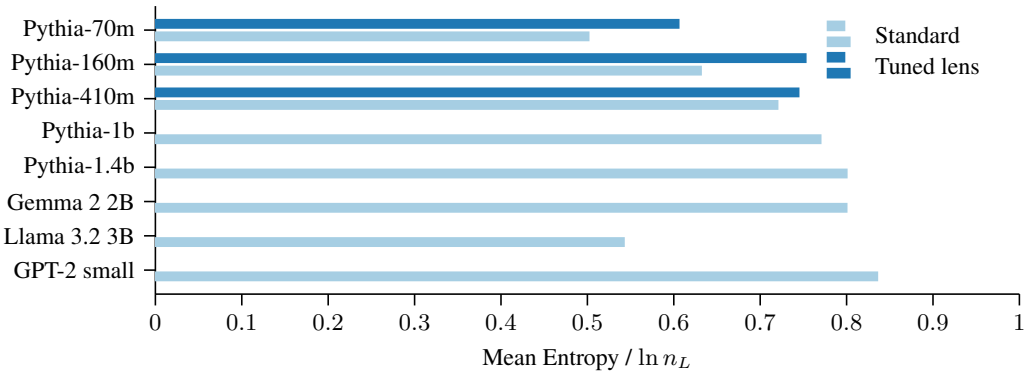


(a) Varying the model with an expansion factor of $R = 64$ and sparsity $k = 32$

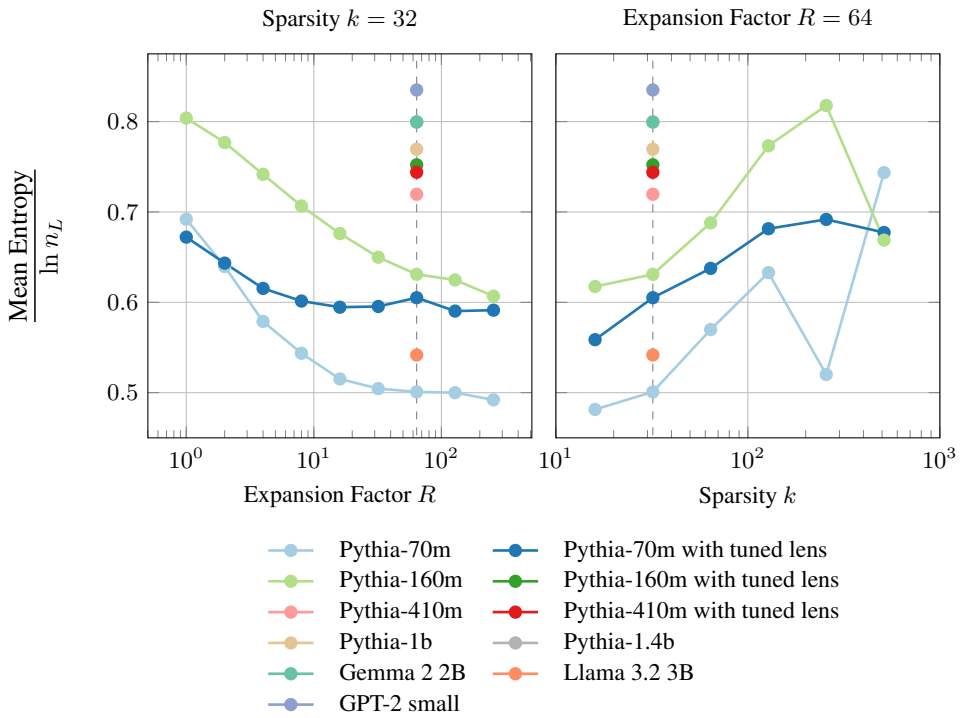


(b) Varying the expansion factor R with sparsity $k = 32$ and k with $R = 64$

Figure 28: The mean number of layers at which latents have a count of non-zero activations above a threshold divided by the total number of model layers, over 10 million tokens from the test set. The threshold is 10 thousand tokens (0.1%). As in Figure 5, the absence of bars for tuned-lens MLSAEs indicates the absence of results, not that the values are zero. The mean active layers decreases as the expansion factor R increases, and increases as the sparsity k increases up to a point. There is no clear trend with respect to the model size. Notably, applying tuned-lens transformations to the input activations from each layer decreases the mean active layers (Section 3.3).



(a) Varying the model with an expansion factor of $R = 64$ and sparsity $k = 32$



(b) Varying the expansion factor R with sparsity $k = 32$ and k with $R = 64$

Figure 29: The mean entropy of the observed discrete distributions of latent activations over layers (Eq. 10) divided by the maximum entropy $\ln n_L$, over 10 million tokens from the test set. As in Figure 5, the absence of bars for tuned-lens MLSAEs indicates the absence of results, not that the values are zero. The entropy tends to decrease as the expansion factor R increases, to increase as the sparsity k increases, and to increase as the model size increases for Pythia transformers.

We computed the entropy of the observed distributions of activations over layers, took the mean over latents, and divided it by $\ln n_L$ to compare models with different numbers of layers. The normalized mean entropy increases slightly as the model size increases for Pythia models (Figure 29a), similarly to the variance of the layer index (Section 4.3). However, it decreases as the number of latents increases relative to the model dimension, similarly to the mean active layers (Figure 29b).

D ADDITIONAL HEATMAPS

For completeness, we include equivalent aggregate and single-prompt heatmaps to Figures 2 and 3 for different models and combinations of hyperparameters:

- Varying R for Pythia-70m and $k = 32$ (Figures 30 and 31)
- Varying k for Pythia-70m and $R = 64$ (Figures 32 and 33)
- Varying R for Pythia-160m and $k = 32$ (Figures 34 and 35)
- Varying k for Pythia-160m and $R = 64$ (Figures 36 and 37)
- Varying R for Pythia-70m with tuned lens and $k = 32$ (Figures 38 and 39)
- Varying k for Pythia-70m with tuned lens and $R = 64$ (Figures 40 and 41)