C Measures of latents active at multiple layers

C.1 Heatmap normalization

In the aggregate and single-prompt heatmaps such as Figures 2 and 3, we plot the distributions of latent activations over layers, which we take to be proportional to the total activations when aggregating over a large sample of tokens (Eq. 10). We normalized the latent activations in this way to visually compare the aggregate and single-prompt heatmaps, as well as individual latents within a heatmap, due to the wide range of activation counts and totals across latents.

Normalizing the activations discards the relative frequencies and magnitudes of activations for different latents, so we reproduce Figures 2 and 3 with the un-normalized totals of latent activations in Figures 25 and 26. We use power-law normalization for the colormaps, i.e., $y = x^\gamma$, to account for the wide range of values; all other heatmaps have linear colormaps. As with all other single-prompt heatmaps, we exclude latents from Figure 26 that never activate. In both cases, the un-normalized results are qualitatively similar to the normalized results.

C.2 Variance of the layer index

Recall that we consider the layer L, token T, and latent index J as random variables (Section 4.3). For a single latent, we have, by the law of total variance:

$$\mathrm{Var}\left[L\right] = \mathbb{E}[\mathrm{Var}\left[L \mid T\right]] + \mathrm{Var}\left[\mathbb{E}[L \mid T]\right] \tag{14}$$

We are interested in the first two terms:

- $\mathrm{Var}\left[L\right]$ is the variance of the distribution over layers, aggregating over tokens;
- $\mathbb{E}[\mathrm{Var}\left[L \mid T\right]]$ is the mean variance of the distributions over layers for each token; and
- $\mathrm{Var}\left[\mathbb{E}[L \mid T]\right]$ is the variance of the mean layers for each token.

Aggregating over latents, we have:

$$\mathbb{E}[\mathrm{Var}\left[L \mid J\right]] = \mathbb{E}[\mathrm{Var}\left[L \mid T, J\right]] + \mathbb{E}[\mathrm{Var}\left[\mathbb{E}[L \mid T, J] \mid J\right]] \tag{15}$$

C.3 Number of layers above a threshold

The count of layers at which a latent is non-zero ('active') does not necessarily positively correlate with the variance of the layer index considered in Section 4.3. For example, the variance of 0 and 5 (two distinct values) is greater than the variance of 2, 3, and 4 (three distinct values). Strictly speaking, the layer index is ordinal data, but we implicitly treat it as interval data by taking the arithmetic mean and variance. We chose this approach because we expected latents to be active over a contiguous range of layers, which is validated qualitatively by the heatmaps such as Figures 2 and 3.

For comparison, we computed the number of layers at which each latent has a count of non-zero activations above a threshold ('active layers') divided by the total number of model layers n_L. We selected a threshold count of 10k tokens (0.1% of a sample of 10M tokens). When aggregating over latents, the relative mean active layers decreases as the model size increases for Pythia models (Figure 28a) and as the number of latents increases relative to the model dimension (Figure 28b). Importantly, this measure depends strongly on the choice of threshold, unlike our variance ratios.

C.4 Entropy

A further measure of the degree to which a latent is active at multiple layers is the statistical distance between the observed discrete distribution of activations over layers (Eq. 10) and a reference distribution. At one extreme is a Dirac distribution with probability mass 1 for a single layer index and 0 elsewhere, in which case the latent is active at a single layer. The other extreme is the discrete uniform distribution $\mathcal{U}(0, n_L)$, in which case the latent is equally active at every layer. Hence, the entropy of the observed distribution ranges between 0 and $\ln n_L$. Notably, the entropy of the observed distribution is agnostic with respect to the numeric values of the layer indices and their ordering.

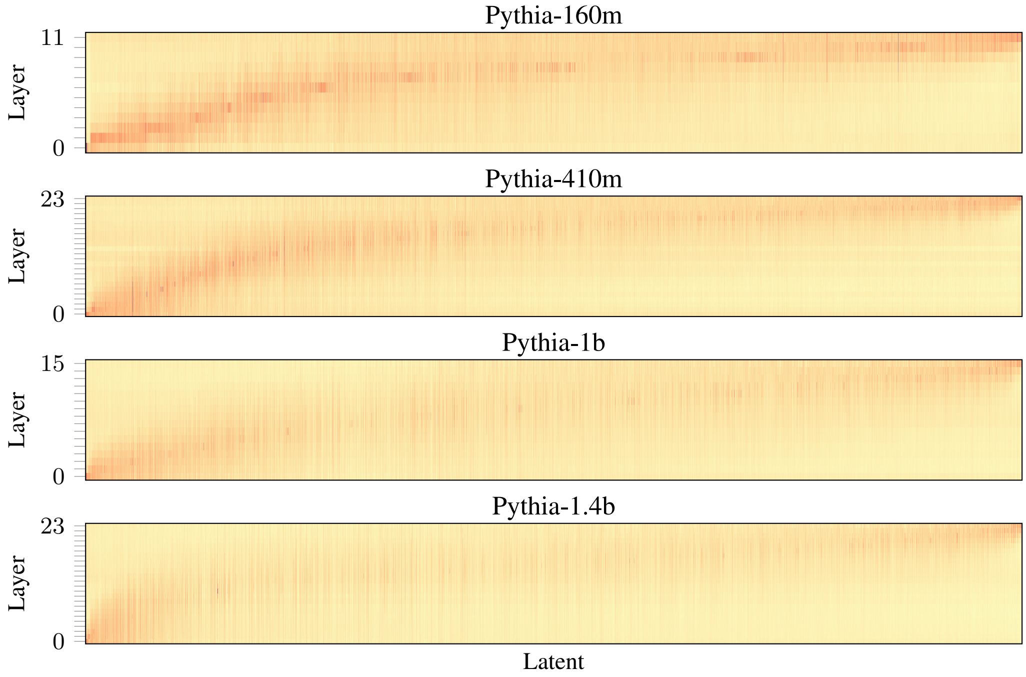

Figure 25: Heatmaps of the total latent activations over layers when aggregating over 10 million tokens from the test set. Here, we plot the totals for MLSAEs trained on Pythia models with an expansion factor of $R = 64$ and sparsity $k = 32$. We provide further details in Figure 2. The colormaps use power-law normalization with $\gamma = 1/4$.

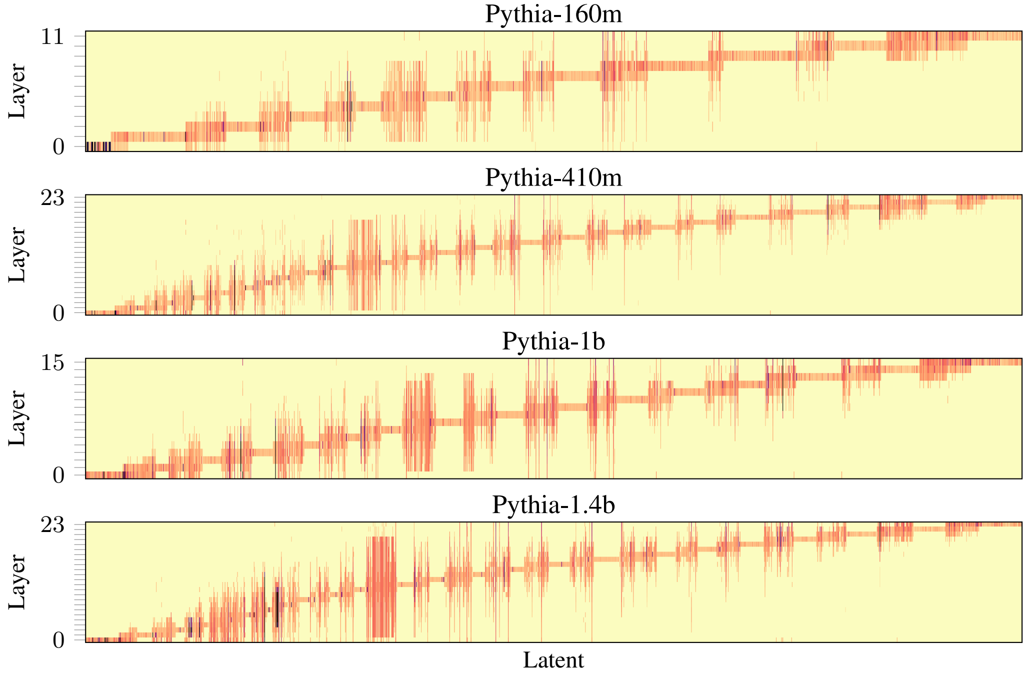

Figure 26: Heatmaps of the total latent activations over layers for a single example prompt. Here, we plot the totals for MLSAEs with an expansion factor of $R = 64$ and sparsity $k = 32$. The example prompt is "When John and Mary went to the store, John gave" (Wang et al., 2022). We provide further details in Figure 3. The colormaps use power-law normalization with $\gamma = 1/2$.

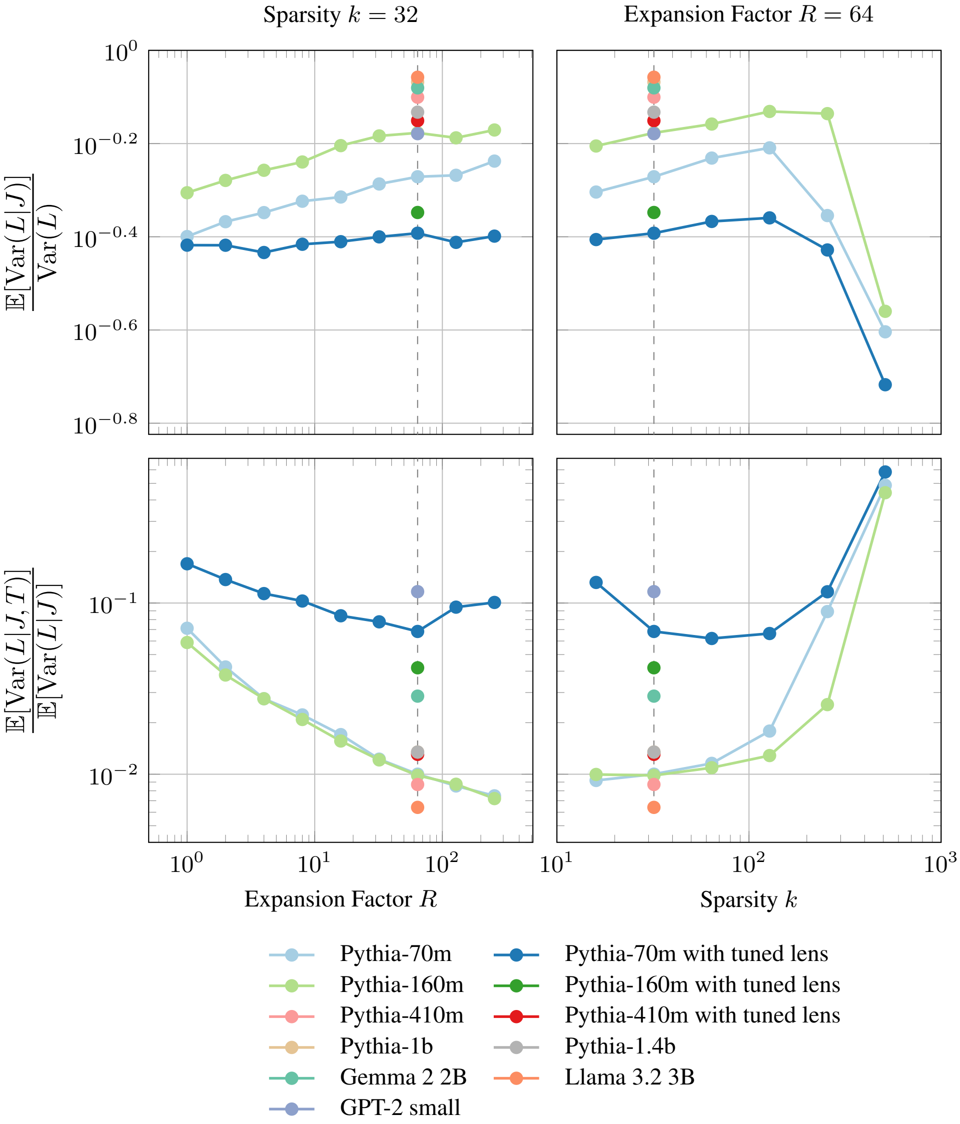

Figure 27: The fraction of the total variance explained by individual latents and the fraction of the variance for an individual latent explained by individual tokens (Eqs. 11 and 12). Here, we plot the variance ratios for standard and tuned-lens MLSAEs over 10 million tokens from the test set.