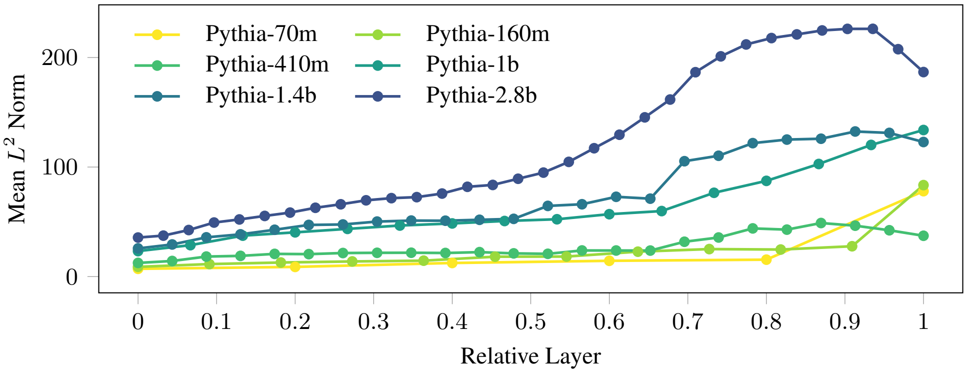

Figure 4: The mean L^2 norm of the residual stream activation vectors at every layer, over 10 million tokens from the test set. To compare transformers with different numbers of layers, we divide the layer index ℓ by the number of layers n_L. This 'relative layer' is the x-axis of the plot.

Following Heimersheim & Turner (2023), we verified that the mean L^2 norm of the activation vectors increases across layers, which prompted us to center the vectors at each layer by subtracting the dataset mean before computing the similarities between vectors (Figure 4).

4.3 Latent distributions over layers

Given a dataset and MLSAE, each combination of a token and latent produces a distribution of activations over layers. We want to understand the degree to which the variance of that distribution depends on the token versus the latent to quantify the intuition gleaned from Figures 2 and 3.

Consider the layer index L, token T, and latent index J to be random variables. We take $P(J)$ to be a uniform discrete distribution, $P(T \mid J)$ to be a uniform discrete distribution over tokens for which the latent is active (at any layer), and L to be sampled from a conditional distribution proportional to the total latent activation at that layer, aggregating over tokens:

$$P(L = \ell \mid T = t,\, J = j) = \frac{h_j(\mathbf{x}_{t,\ell})}{\sum_{\ell'} h_j(\mathbf{x}_{t,\ell'})} \tag{10}$$

Here, $\mathbf{x}_{t,\ell}$ is the dense residual stream activation vector at token t and layer ℓ, while $h_j(\mathbf{x}_{t,\ell})$ is the activation of the j-th MLSAE latent at that token and layer.

We order latents in all heatmaps using the expected value of the layer index for a single latent $\mathbb{E}[L \mid J = j]$. The variance of the distribution over layers measures the degree to which a latent is active at a single layer (in which case, it is zero) versus multiple layers (in which case, it is positive). We are interested in the following variances of the distribution over layers:

- $\mathrm{Var}[L \mid J = j,\, T = t]$, for a single latent and token
- $\mathrm{Var}[L \mid J = j]$, for a single latent, aggregating over tokens
- $\mathrm{Var}[L]$, aggregating over both latents and tokens

These quantities are related by the law of total variance (see Appendix C.2). For the moment, we note that the variance of the distribution over layers naturally depends on the number of layers n_L. Hence, to compare different models, we look at ratios between these variances:

$$\begin{array}{c}\text{Variance for one latent, aggregating over tokens,}\\ \text{as a proportion of the total variance over all latents}\end{array} = \frac{\mathbb{E}[\mathrm{Var}(L \mid J)]}{\mathrm{Var}(L)} \tag{11}$$

$$\begin{array}{c}\text{Variance for one token and latent as a}\\ \text{proportion of the total variance for that latent}\end{array} = \frac{\mathbb{E}[\mathrm{Var}(L \mid J,\, T)]}{\mathbb{E}[\mathrm{Var}(L \mid J)]} \tag{12}$$

The former measures the degree to which latents are active at multiple layers when aggregating over tokens, and the latter compares this to the case for a single token. We explore alternative measures of aggregate multi-layer activity in Appendix C.

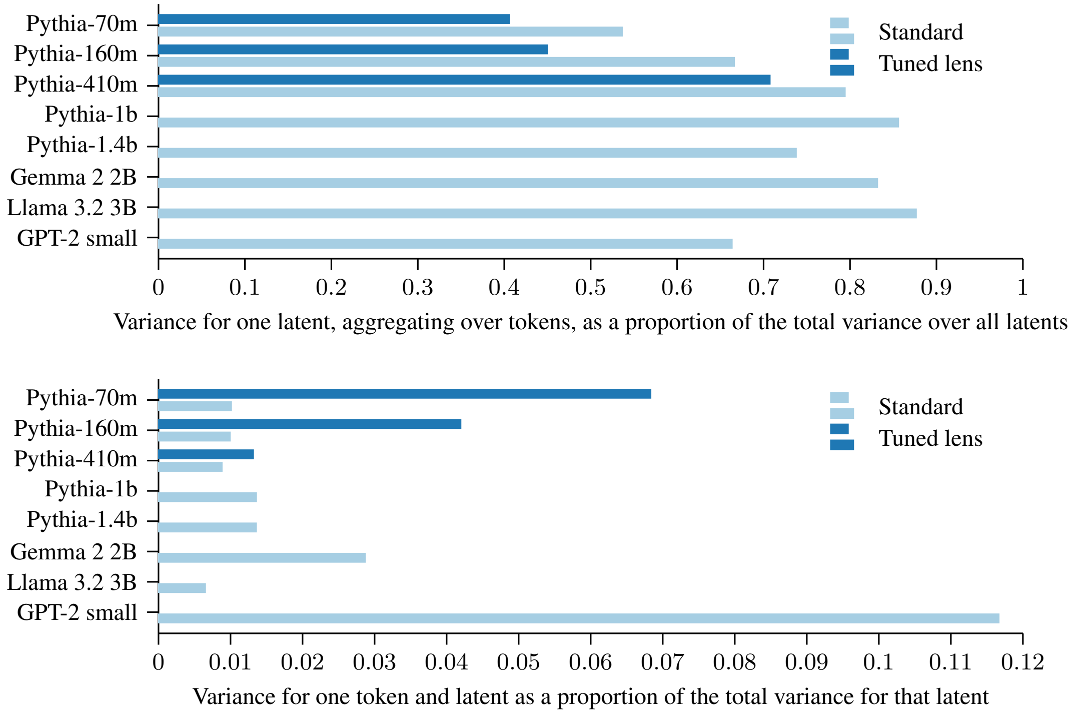

Figure 5: The fraction of the total variance explained by individual latents and the fraction of the variance for an individual latent explained by individual tokens (Eqs. 11 and 12) for MLSAEs with an expansion factor of $R = 64$ and sparsity $k = 32$, over 10 million tokens from the test set. The absence of bars for tuned-lens MLSAEs indicates the absence of results, not that the values are zero.

The degree to which latents are active at multiple layers when aggregating over tokens is relatively large, between 54 and 88%, and broadly increases with the model size for fixed hyperparameters (Figure 5). This measure quantifies the observation that, in the aggregate heatmaps (Figure 2), the distributions of latent activations over layers become more ‘spread out’ as the model size increases. Conversely, we find that the fraction of the variance for an individual latent explained by individual tokens is relatively small, on the order of 1 to 10%. This quantifies the observation that, in the single-prompt heatmaps (Figure 3), the distributions over layers are much less ‘spread out’ than in the aggregate heatmaps.

4.4 Tuned lens

Thus far, we have assumed that the residual stream basis is the same at every layer. We relaxed this assumption by applying pre-trained tuned-lens transformations to the residual stream activations at each layer before the encoder (Section 3.3). We had expected that these transformations would increase the degree to which latents were active at multiple layers because they translate the activations at every layer into a basis more similar to the basis of the output layer. The aggregate and single-prompt heatmaps (Figures 6 and 7) indicate a modest increase in the degree to which latents are active at multiple layers compared with the standard approach.

The variance ratios in Figure 5 clarify that the tuned-lens approach decreases the degree to which latents are active at multiple layers when aggregating over tokens. This ratio remains approximately constant as the expansion factor increases, between 37% and 41% (Figure 27). Conversely, the variances for a single token relative to a single latent are larger, i.e., the single-prompt heatmaps are more ‘spread out’ compared with the standard approach.

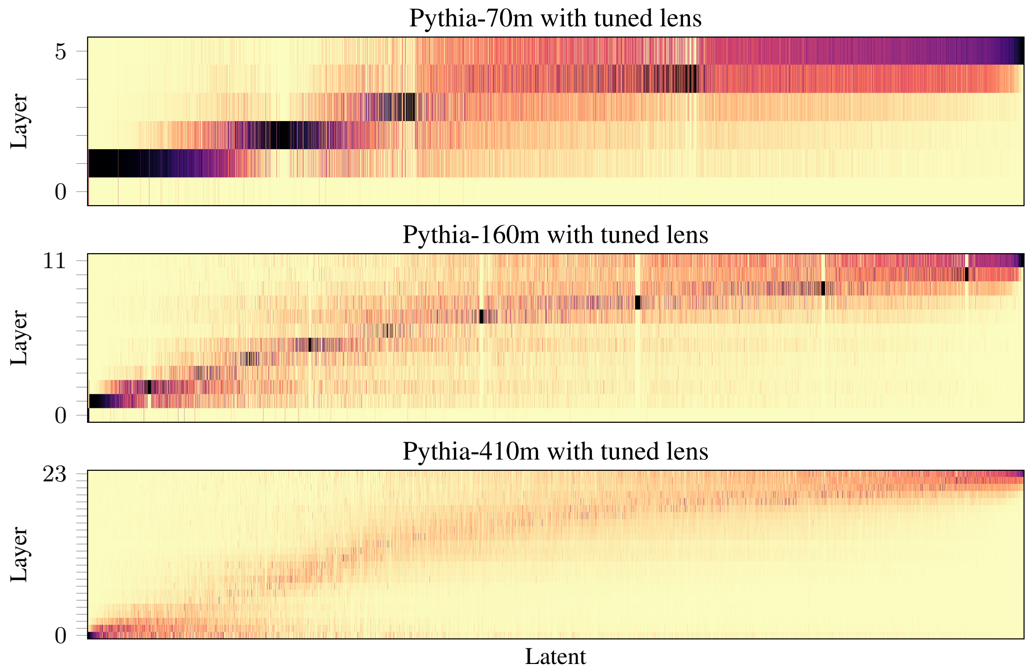

Figure 6: Heatmaps of the distributions of latent activations over layers when aggregating over 10 million tokens from the test set. Here, we plot the distributions for tuned-lens MLSAEs trained on Pythia models with an expansion factor of $R = 64$ and sparsity $k = 32$. For standard MLSAEs, see Figure 2. We note that a pre-trained tuned lens was not available for Pythia-1b (Section 3.3).

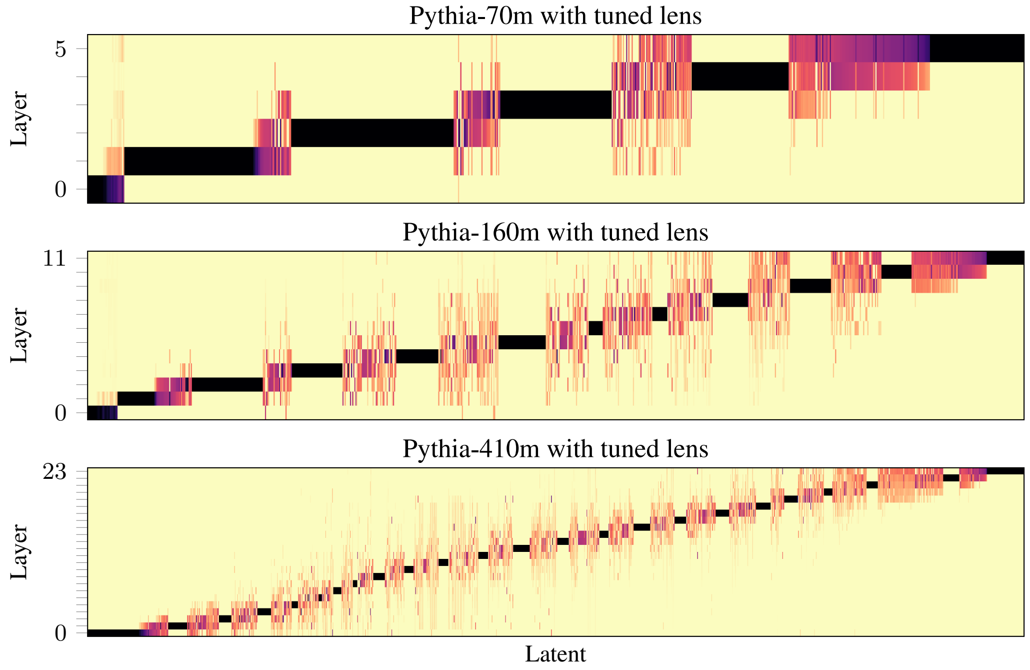

Figure 7: Heatmaps of the distributions of latent activations over layers for a single example prompt. Here, we plot the distributions for tuned-lens MLSAEs trained on Pythia models with an expansion factor of $R = 64$ and sparsity $k = 32$. The example prompt is “When John and Mary went to the store, John gave” (Wang et al., 2022). For standard MLSAEs, see Figure 3. We note that a pre-trained tuned lens was not available for Pythia-1b (Section 3.3).