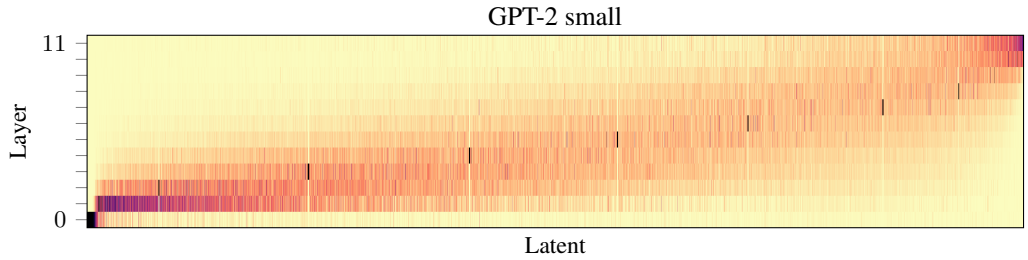

Figure 13: Heatmaps of the distributions of latent activations over layers when aggregating over 10 million tokens from the test set. Here, we plot the distributions for MLSAEs trained on GPT-2 small with an expansion factor of $R = 64$. We provide further details in Figure 2.

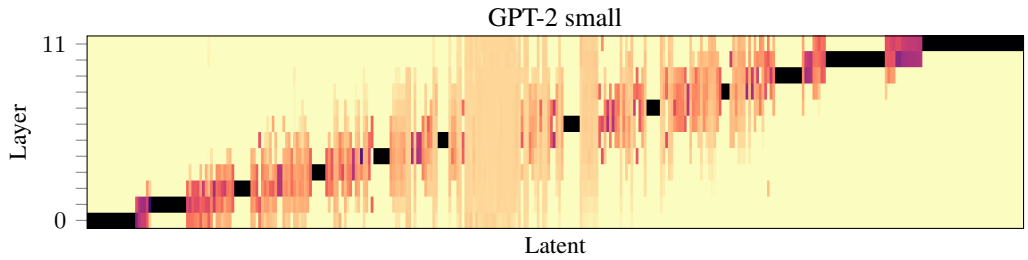

Figure 14: Heatmaps of the distributions of latent activations over layers for a single example prompt. Here, we plot the distributions for MLSAEs trained on GPT-2 small with an expansion factor of $R = 64$. The example prompt is "When John and Mary went to the store, John gave" (Wang et al., 2022). We provide further details in Figure 3.

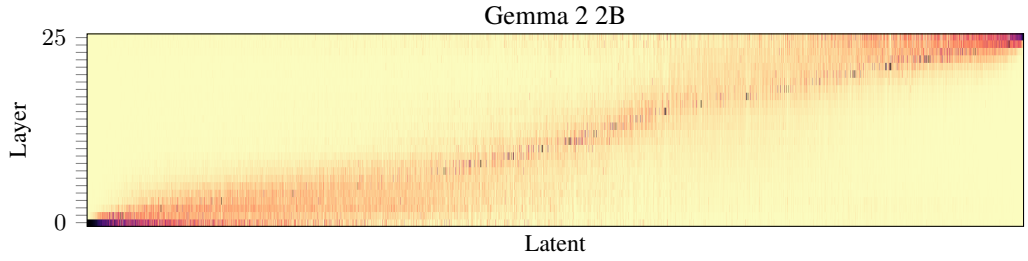

Figure 15: Heatmaps of the distributions of latent activations over layers when aggregating over 10 million tokens from the test set. Here, we plot the distributions for MLSAEs trained on Gemma 2 2B with an expansion factor of $R = 64$. We provide further details in Figure 2.

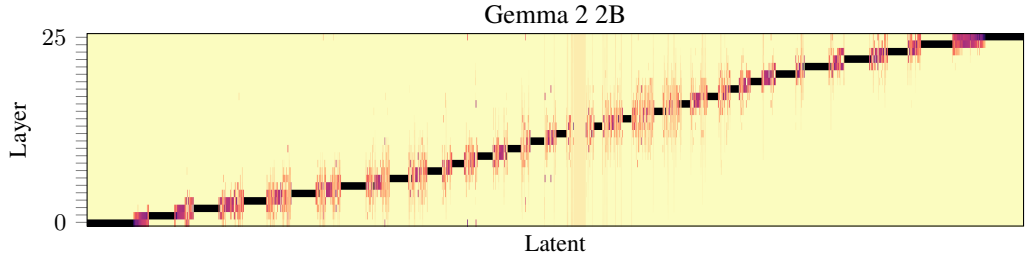

Figure 16: Heatmaps of the distributions of latent activations over layers for a single example prompt. Here, we plot the distributions for MLSAEs trained on Gemma 2 2B with an expansion factor of $R = 64$. The example prompt is "When John and Mary went to the store, John gave" (Wang et al., 2022). We provide further details in Figure 3.

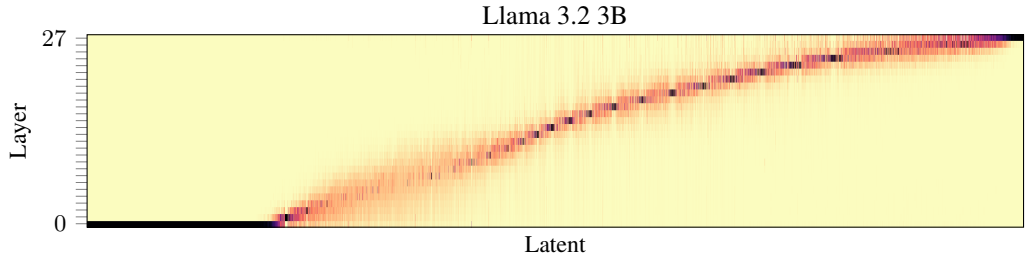

Figure 17: Heatmaps of the distributions of latent activations over layers when aggregating over 10 million tokens from the test set. Here, we plot the distributions for MLSAEs trained on Llama 3.2 3B with an expansion factor of $R = 64$. We provide further details in Figure 2. Notably, a greater proportion of latents are only active at the first layer compared with other transformer architectures.

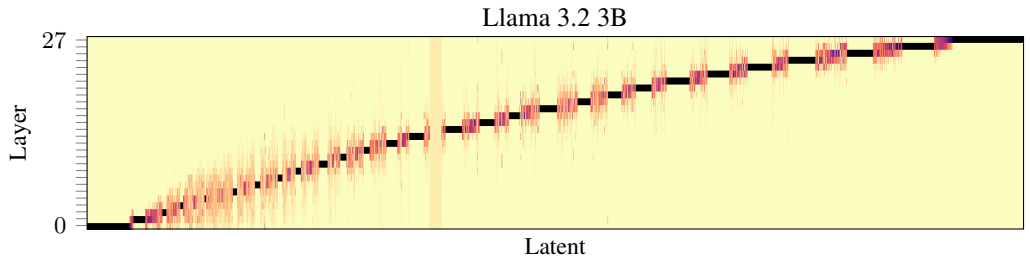

Figure 18: Heatmaps of the distributions of latent activations over layers for a single example prompt. Here, we plot the distributions for MLSAEs trained on Llama 3.2 3B with an expansion factor of $R = 64$. The example prompt is "When John and Mary went to the store, John gave" (Wang et al., 2022). We provide further details in Figure 3.

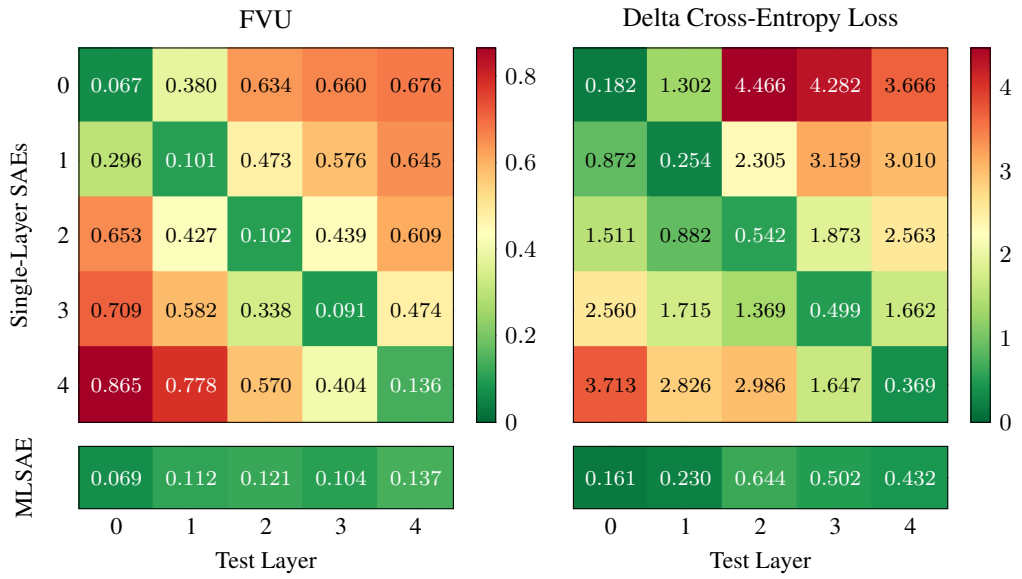

Figure 19: The FVU reconstruction error and delta cross-entropy loss for single-layer SAEs trained on each layer of Pythia-70m, compared with a single multi-layer SAE trained on every layer. The single-layer SAEs trained on data from a given layer perform best on test data from the same layer (the diagonal elements of the matrix plot); a multi-layer SAE trained on data from every layer performs comparably to the corresponding single-layer SAEs (the row beneath the matrix plot).

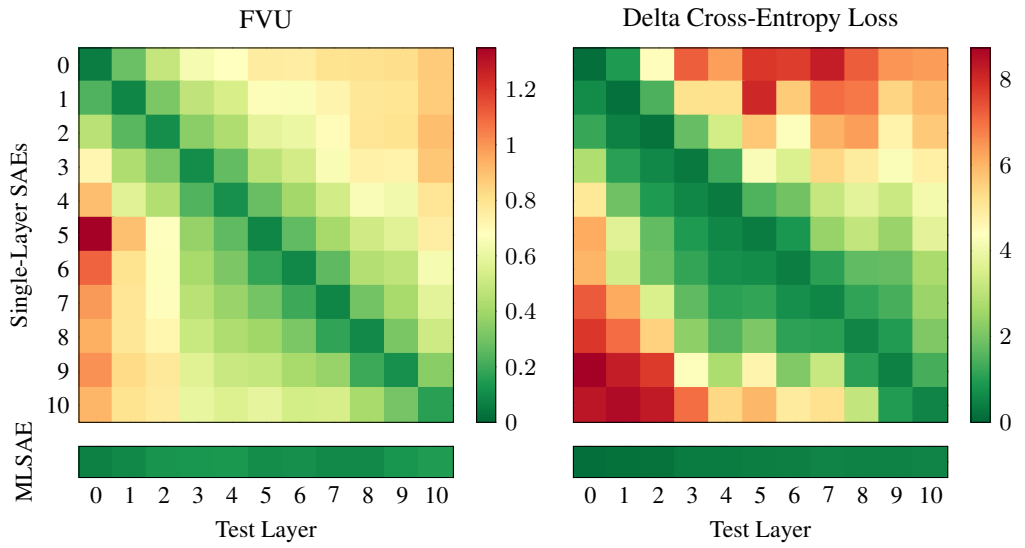

Figure 20: The FVU reconstruction error and delta cross-entropy loss for single-layer SAEs trained on each layer of Pythia-160m, compared with a single multi-layer SAE trained on every layer. The single-layer SAEs trained on data from a given layer perform best on test data from the same layer (the diagonal elements of the matrix plot); a multi-layer SAE trained on data from every layer performs comparably to the corresponding single-layer SAEs (the row beneath the matrix plot).

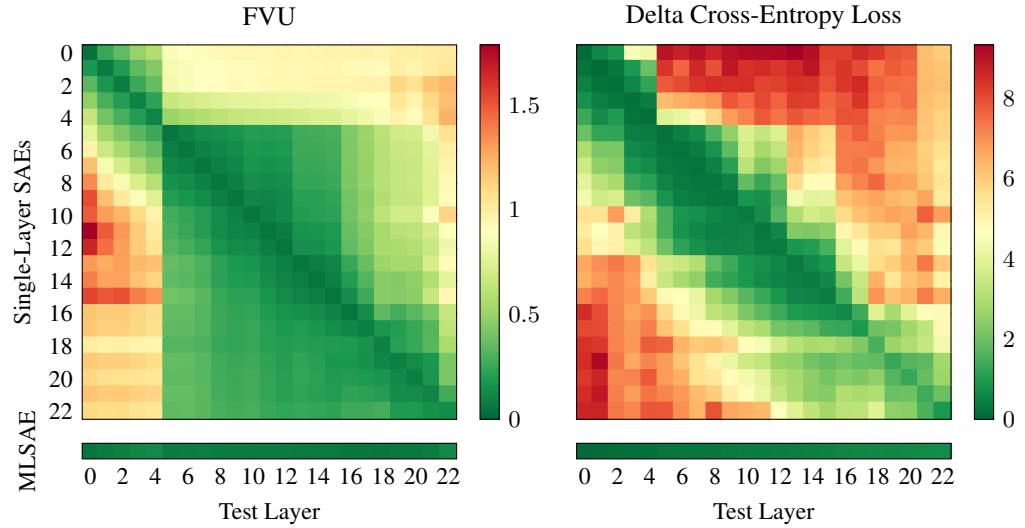

Figure 21: The FVU reconstruction error and delta cross-entropy loss for single-layer SAEs trained on each layer of Pythia-410m, compared with a single multi-layer SAE trained on every layer. The single-layer SAEs trained on data from a given layer perform best on test data from the same layer (the diagonal elements of the matrix plot); a multi-layer SAE trained on data from every layer performs comparably to the corresponding single-layer SAEs (the row beneath the matrix plot). Interestingly, there is a sharp change in the FVU between single-layer SAEs trained on layers 4 and 5.