Figure 39: 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-70m with 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.
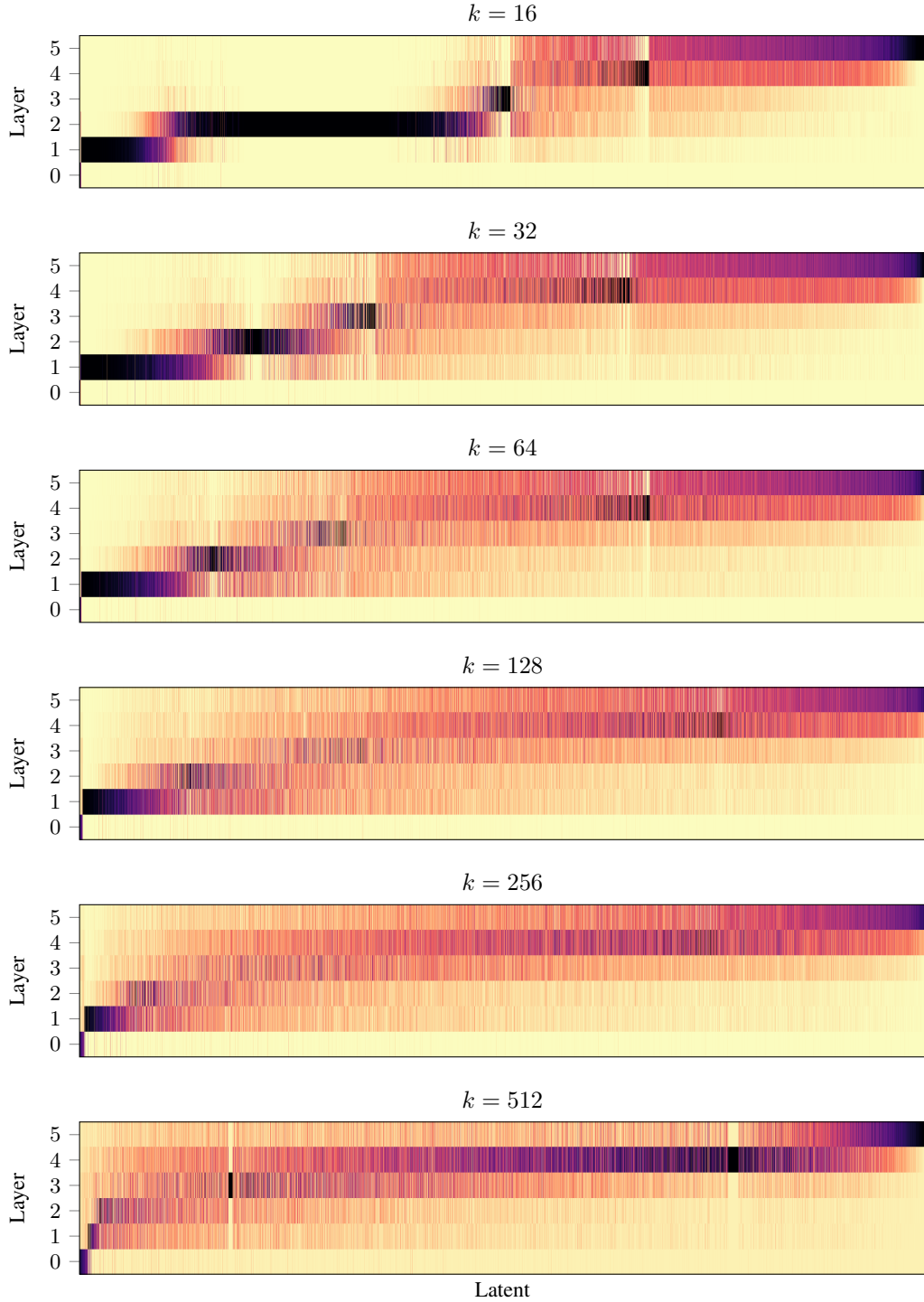
Figure 40: Heatmaps of the distributions of latent activations over layers when aggregating over 10 million tokens from the test set. Here, we plot the distributions for tuned-lens MLSAEs trained on Pythia-70m with an expansion factor of $R = 64$. We provide further details in Figure 2.
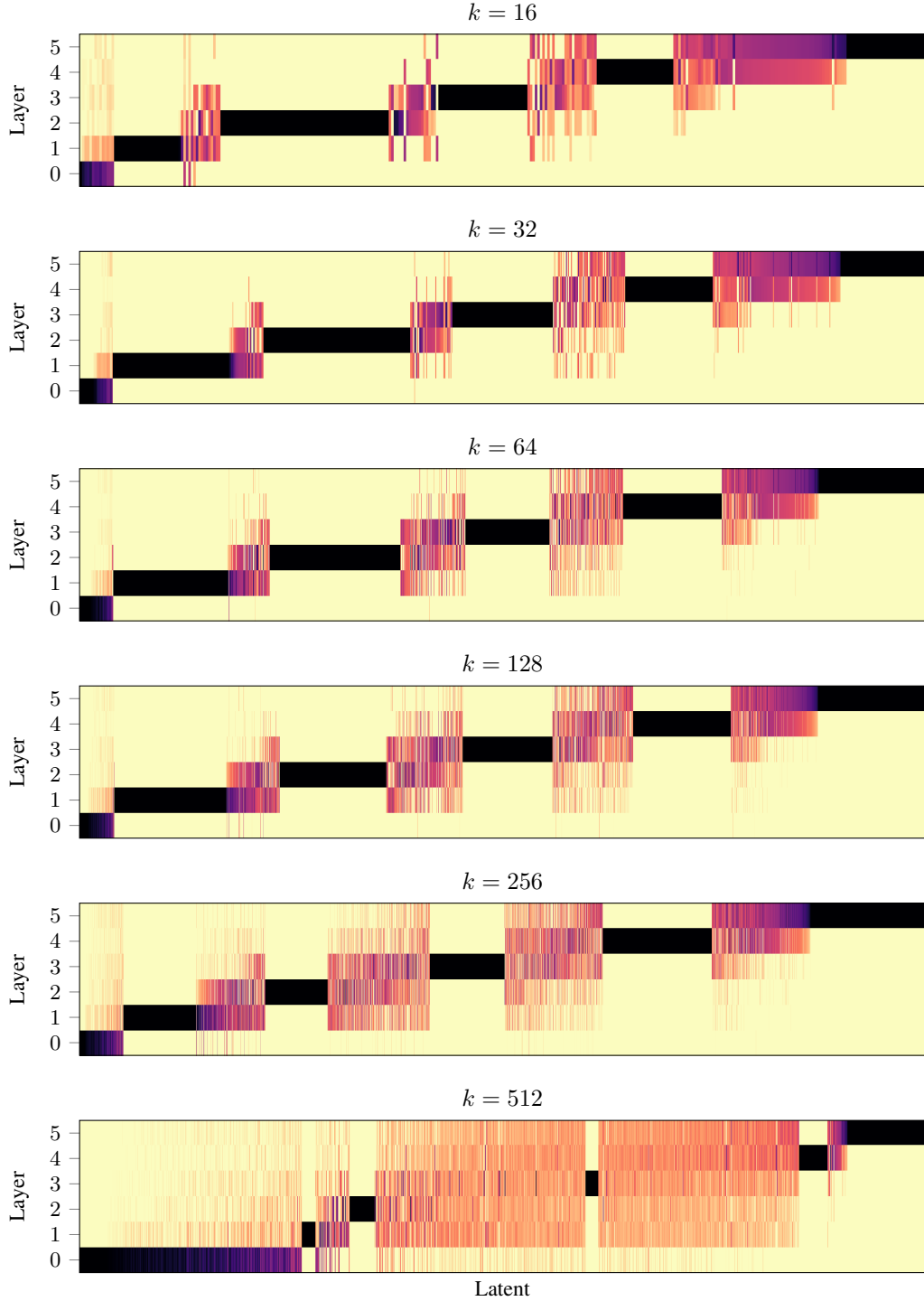
Figure 41: Heatmaps of the distributions of latent activations over layers for a single example prompt. Here, we plot the distributions for tuned-lens MLSAEs trained on Pythia-70m with an expansion factor of $R = 64$. The example prompt is "When John and Mary went to the store, John gave" (Wang et al., 2022). We provide further details in Figure 3.