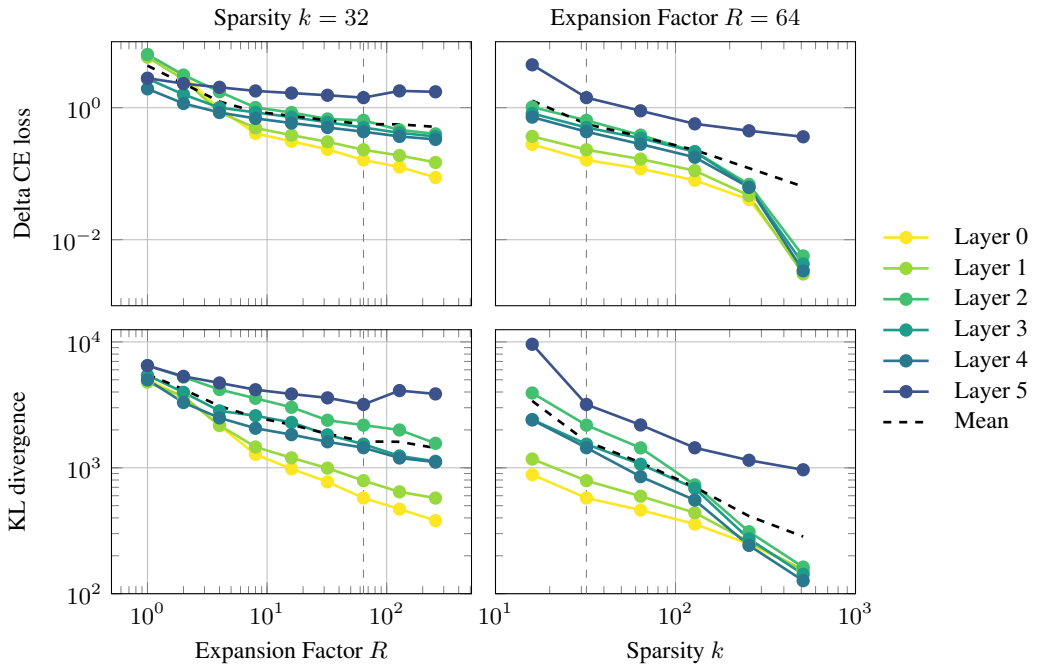
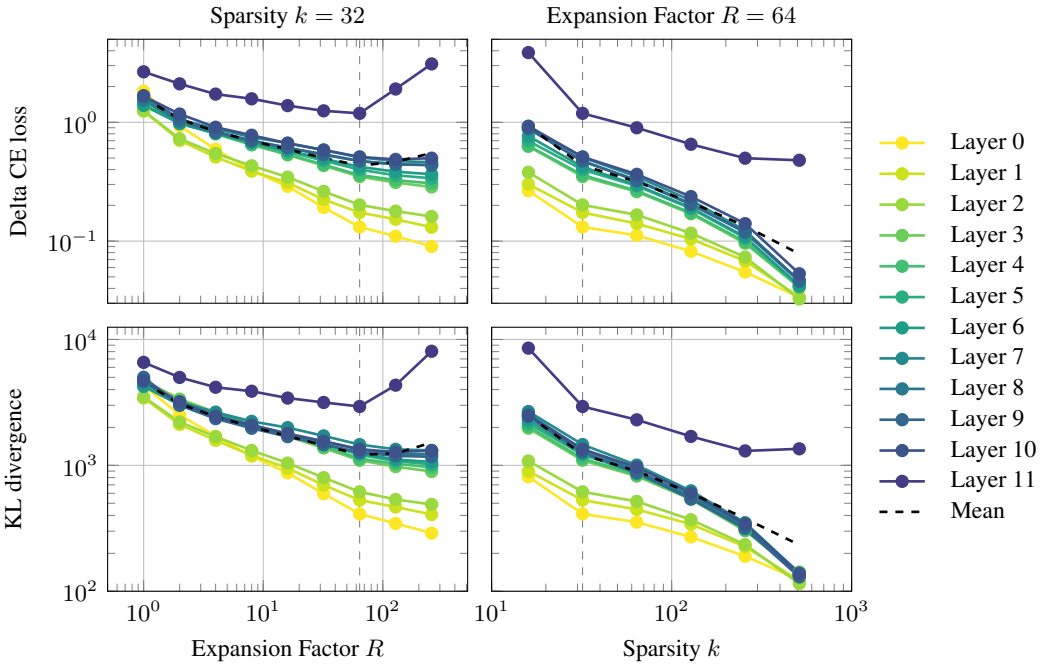


(a) Pythia-70m



(b) Pythia-160m

Figure 10: With fixed sparsity $k = 32$, the delta CE loss and KL divergence generally decrease as the expansion factor increases, except for inputs from the last layer. With fixed expansion factor $R = 64$, both metrics decrease as the sparsity k increases, similarly to the FVU and MSE (Figure 8).

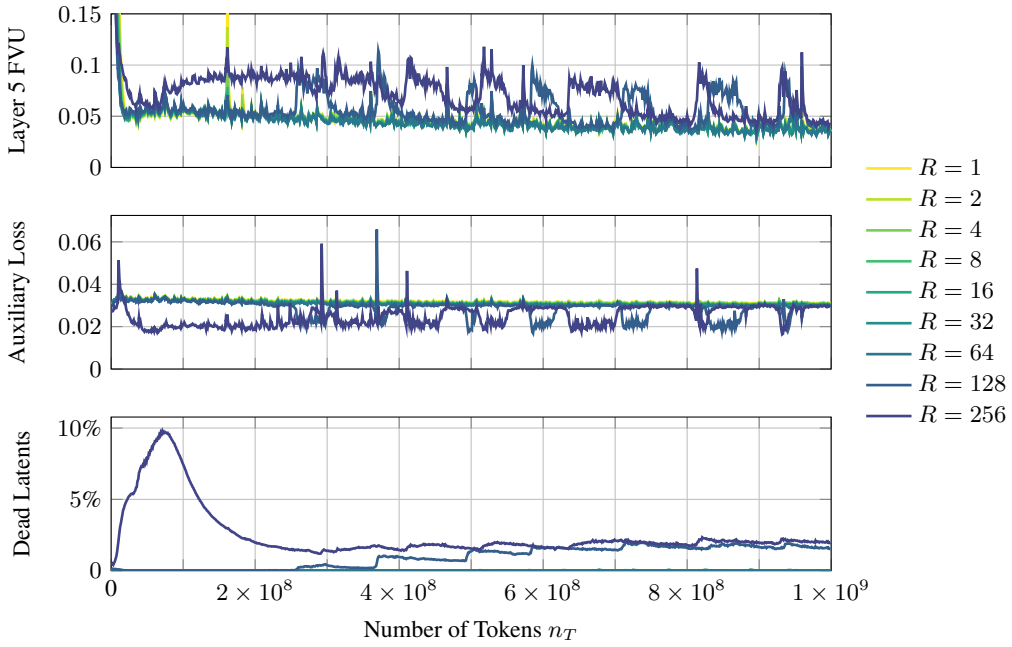


Figure 11: An illustration of the FVU for inputs from the last layer, compared to the auxiliary loss and percentage of dead latents, for MLSAEs trained on Pythia-70m with fixed sparsity $k = 32$. An increase in dead latents correlates with a decrease in the auxiliary loss and an increase in the FVU at the last layer. We attribute this to the increased scale of the inputs because the auxiliary loss depends on the MSE (Figure 8). The auxiliary loss is multiplied by its coefficient $\alpha = 1/32$ in the training loss.

We include quantitative results for Gemma 2 2B, Llama 3.2 3B, and GPT-2 small in Table 1, Figure 5, and throughout Appendix B. We include heatmaps of the distributions of latent activations over layers in Figures 13, 14, 15, 16, 17, and 18, which are qualitatively similar to the Pythia models. Interestingly, we observed that our validation metrics improved at a slower rate for Gemma 2 2B, so we continued to train this MLSAE up to a total of approximately 2.5 billion tokens. Nevertheless, the increase in the cross-entropy loss remains larger than other models of similar sizes (Table 1).

B.4 Single-layer SAEs

While we compare the performance of our multi-layer SAEs to single-layer SAEs from the literature in Appendix B.1 and B.2, we also trained multiple single-layer SAEs on Pythia-70m, 160m, and 410m with our default hyperparameters, leaving the remainder of the experimental setup unchanged.

Predictably, we find that a single-layer SAE trained on data from a given layer performs best on test data from the same layer (Figures 19, 20, and 21). A multi-layer SAE trained on data from every layer performs comparably to the corresponding single-layer SAE, and more consistently across test data from different layers. Interestingly, applying the corresponding tuned-lens transformation to the input activations from each layer during training and evaluation degrades the performance of single-layer SAEs on test data from different layers of Pythia-70m, unlike multi-layer SAEs (Figure 12).

Importantly, the results for the last layer are excluded from these figures. This is because we take the residual stream activation vectors after a given layer has been applied (Section 3.1), such that the last-layer activations represent only the next-token predictions of the model and not intermediate computational variables. Hence, we expect these activations to have a significantly different structure to the preceding layers, which could distort comparisons between layers (Lad et al., 2024).

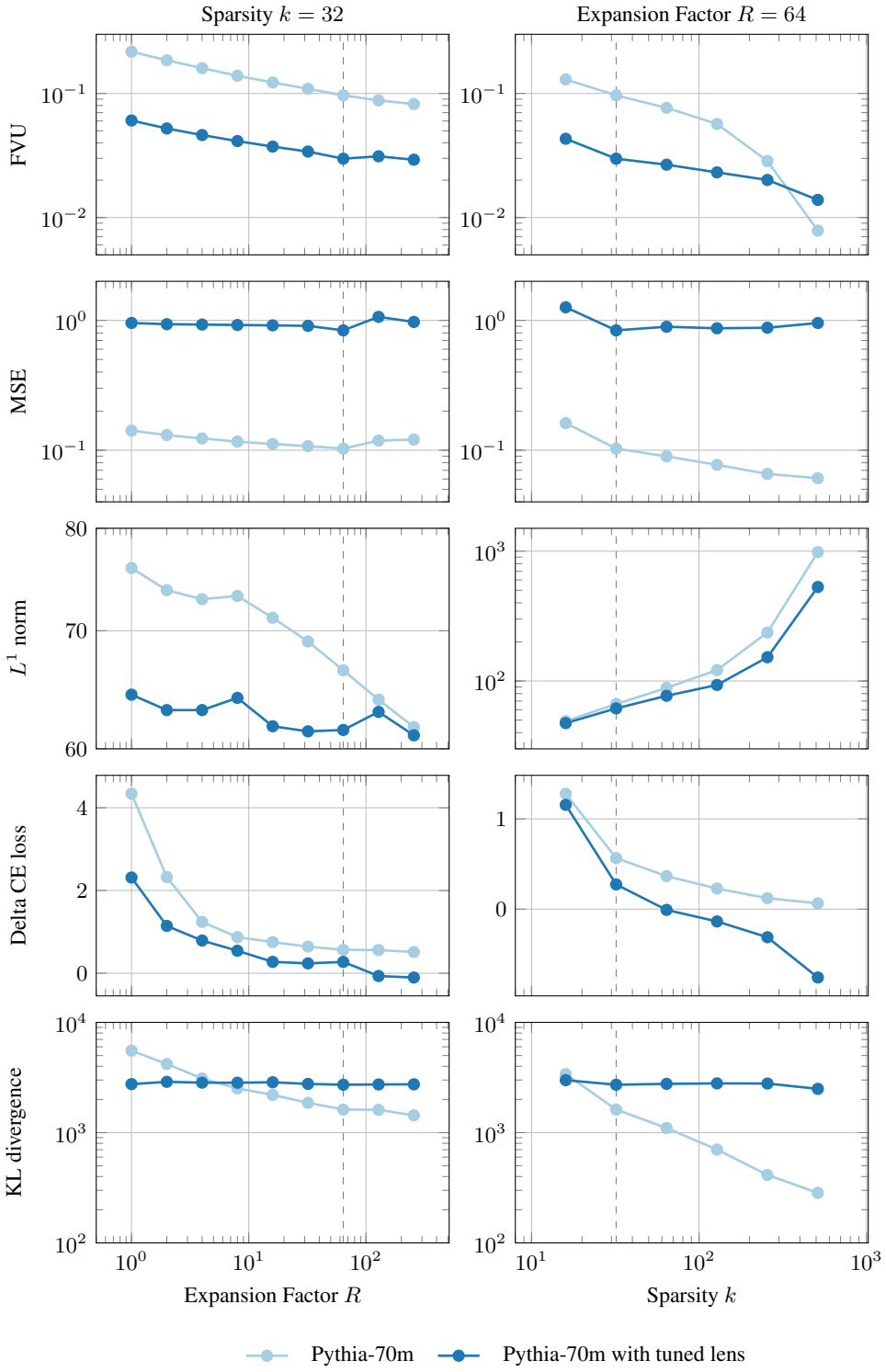


Figure 12: For Pythia-70m, applying tuned-lens transformations decreases the mean FVU and delta cross-entropy loss but not the KL divergence. Importantly, we compute reconstruction errors before applying the inverse transformation and downstream loss metrics afterward (Section 3.3). Unlike Figure 10, we use a linear scale for the delta cross-entropy loss because, surprisingly, it is negative for tuned-lens MLSAEs with a large expansion factor R or sparsity k; see also Figure 22.