G AUTOREGRESSIVE DECODER-ONLY MODELS

In this section, we further validate our sparsity scaling laws in the context of standard decoder-only Transformers trained for auto-regressive language modeling. In general, we follow the same setup as in our T5/C4 experiments, changing only model architecture and loss function. We again execute the experiment grid defined in Table 3 for 250K and 500K training steps, as well as a subset of the more expensive 1M step runs.

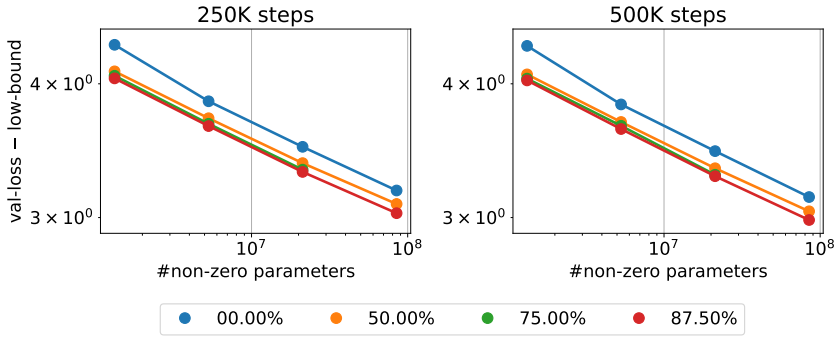

Figure 8: Visualization of decoder-only/C4 sweep results for sizes and sparsities.

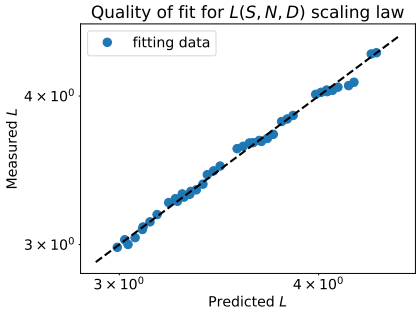

Figure 9: Fit quality of the $L(S, N, D)$ law for decoder-only/C4.

Figure 8 visualizes the collected data, in the same fashion as Section 2.1, demonstrating that all key properties of our scaling law (nearly parallel lines, diminishing returns for high sparsity, near constant loss shifting for more training) can be observed in the context of decoder-only models as well. Consequently, we can also find a good coefficient fit, the result of which is shown in Figure 9.