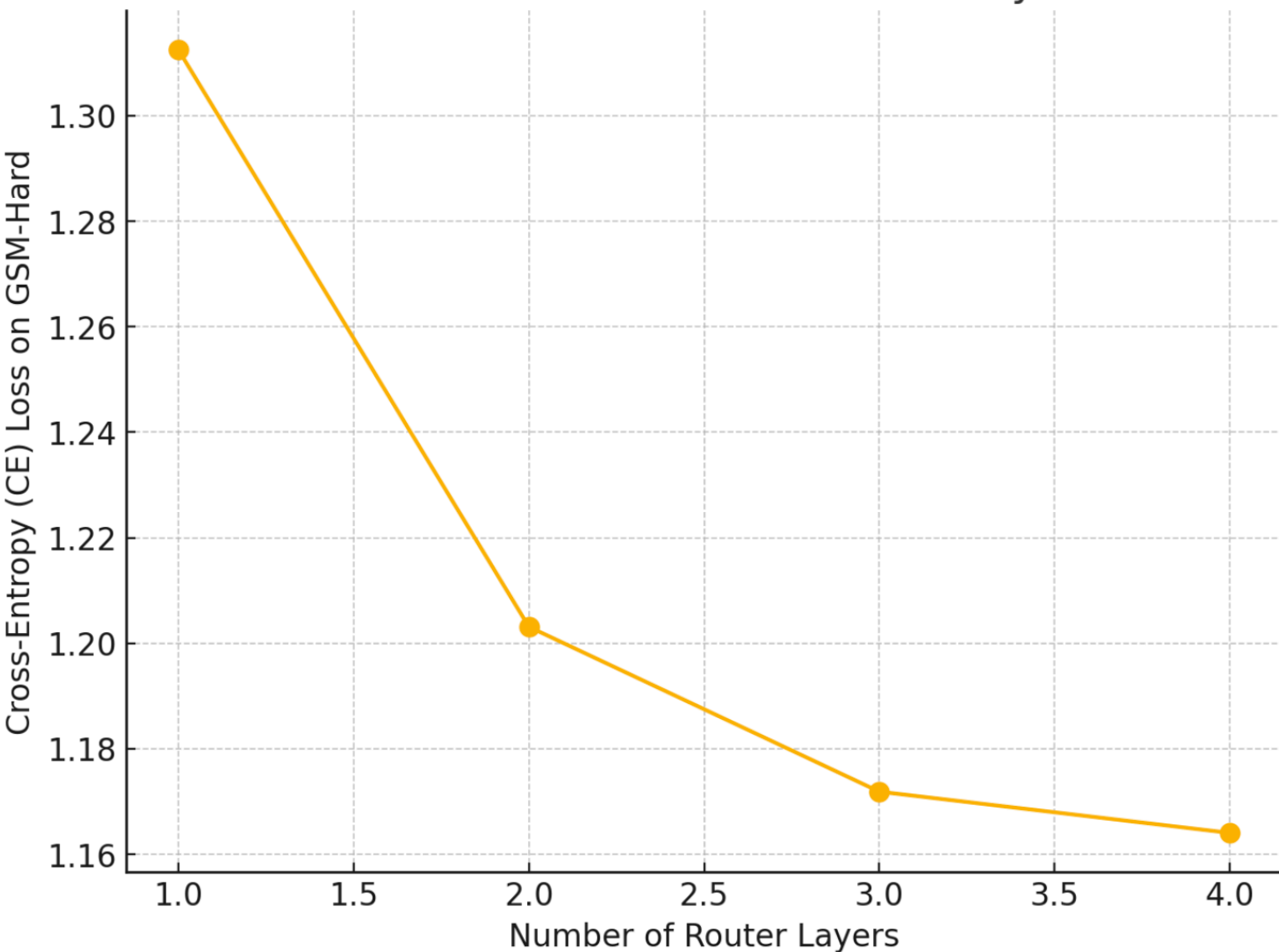


We reran the merging performance vs the representation similarity metric using mutual-knn as our metric. Each dot represents a different time step in the finetuning process, where we use activation interpolation to merge the models. The same “U” shape curve appears.

Test Loss vs. Number of Router Layers



We extended the cross domain (math and coding) to 4 layers as well. The same trend holds where gains are smaller at each increment.