cos similarity heatmap - layer 1 ert-xlarge-v2_uniform-tokens_(Qx,Ky)_trained-params_dense-on_24-layers) 0.75 100 0.50 cosine similarity 0.25 200 token i 0.00 300 -0.25-0.50400 -0.75500 -1.00200 500 100 300 400 token j