cos similarity heatmap - layer 16 ert-xlarge-v2_uniform-tokens_(Qx,Ky)_trained-params_dense-off_24-layers) 0.75 100 -0.50 0.25 cosine similarity 200 token i 0.00 300 -0.25-0.50400 -0.75500 -1.00100 200 500 300 400 token j