Heatmaps of cos similarity $\langle \frac{x_i}{\|x_i\|}, \frac{x_j}{\|x_j\|} \rangle$ (wikitext_albert-xlarge-v2_params-random_dense-on_layers-48, sample 9) layer 25 layer 26 layer 28 layer 27 1.0 1.0 100 100 token itoken itoken i0.0 0.0 0.0 0.0 200 200 -200 200 --0.5-0.5-0.5300 -300 300 --1.0-1.0-1.0100 200 300 100 200 300 100 200 300 100 200 300 token jtoken jtoken jtoken jlayer 29 layer 30 layer 31 layer 32 1.0 1.0 1.0 0 0.5 0.5 0.5token itoken itoken itoken i0.0 0.0 0.0 0.0 200 200 200 -200 -0.5-0.5-0.5300 -300 300 -300 --1.0-1.0-1.0300 200 300 200 300 200 200 300 100 100 100 100 token jtoken \boldsymbol{j} token \boldsymbol{j} token jlayer 36 layer 33 layer 34 layer 35 1.0 1.0 1.0 1.0 0 0 0 token *i*. 100 100 token itoken itoken i0.0 0.0 0.0 0.0 200 200 200 -0.5-0.5300 -300 300 -1.0-1.0200 300 200 300 100 200 300 100 200 300 0 100 100 token \boldsymbol{j} token jtoken \boldsymbol{j} token jlayer 40 layer 38 layer 39 layer 37 1.0 1.0 0 0.5 - 0.5 - 0.5 0.5100 token itoken itoken itoken i0.0 0.0 0.0 0.0200 200 200 -200 -0.5-0.5-0.5300 · 300 300 -300 -1.0-1.0100 200 300 100 200 300 100 200 300 100 200 300 0 token \boldsymbol{j} token \boldsymbol{j} token \boldsymbol{j} token jlayer 41 layer 42 layer 43 layer 44 1.0 1.0 1.0 1.0 0 0 0 token i 0.5 - 0.5 0.5100 100 100 token itoken itoken i0.0 0.0 0.0 200 200 -0.5-0.5300 -300 -300 -300 --1.0-1.0-1.0-1.0100 200 300 100 200 300 100 200 300 100 200 300 token jtoken jtoken jtoken jlayer 45 layer 48 layer 46 layer 47 **-** 1.0 1.0 **-** 1.0 0 . - 0.5 0.5 100 token itoken itoken i0.0 0.0 0.0 200 200 -200 200

300 -

100

200

token \boldsymbol{j}

300

300 -

100

200

 ${\rm token}\ j$

300

300 -

0

100

200

token j

300

300 -

100

200

token j

300