Heatmaps of attention logits $\frac{1}{\sqrt{d'}}\langle Qx_i, Kx_j\rangle$ $({\tt wikitext_albert_xlarge_v2_params_random_dense_off_layers_48})$ layer 26 layer 25 layer 27 layer 28 0 0 0 100 100 100 · 100 token 200 token 5.00 token itoken i200 200 300 300 300 300400 400 400 400 200 400 200 400 200 400 200 400 token jtoken jtoken jtoken \boldsymbol{j} layer 29 layer 30 layer 31 layer 32 0 100 -100 100 100 token 200 token 300 token itoken i200 200 300300 300 300400 400 400 400200 400 200 400 200 400 200 400 token \boldsymbol{j} token jtoken \boldsymbol{j} token \boldsymbol{j} layer 33 layer 34 layer 35 layer 36 0 100 100 100 -100 token 200 token itoken itoken *i*. 200 200 300300 300 300 400 · 400 400 400 200 200 400 200 200 400 400 400 0 0 token jtoken jtoken \boldsymbol{j} token jlayer 40 layer 37 layer 38 layer 39 0 100 -100 · 100 100 token 200 token 500 token itoken i200 200 300 300 300 300400 400 400 · 400200 400 200 400 200 400 200 400 0 token jtoken jtoken \boldsymbol{j} token \boldsymbol{j} layer 41 layer 42 layer 43 layer 44 0 0 0 0 100 100 100 -100 token itoken 200 token itoken i200 200 200 400 -400 -400 -400 -200 200 400 200400 400200 4000 token \boldsymbol{j} token \boldsymbol{j} token jtoken \boldsymbol{j} layer 47 layer 45 layer 46 layer 48 0 100 100 -100 100 token 500 token 300 token *i* 200 -200 300300 300 300 400 · 400 400400 200 200 200 200

400

token \boldsymbol{j}

400

token \boldsymbol{j}

400

token \boldsymbol{j}

400

token \boldsymbol{j}