Heatmaps of attention logits $\frac{1}{\sqrt{d'}}\langle Qx_i, Kx_j\rangle$ $(\verb|imdb_albert-x|| arge-v2_params-trained_dense-off_layers-48)$ layer 27 layer 25 layer 26 layer 28 0 0 - 15 0 15 100 100 100 100 10 10 token 300 300 token 300 300 token 300 · · · · 200 · token 300 300 400 400 400 · 400 500 · 500 $500 \cdot$ 500400 400 200 200400200 200 400 token jtoken \boldsymbol{j} token jtoken jlayer 32 layer 29 layer 30 layer 31 0 0 0 0 100 100 100 -100 - 10 oken 300 300 token 300 300 oken 300 300 token 300 400 400 400 400 500 500 500 500 200 400 200 400 200 400 200 400 token \boldsymbol{j} token jtoken \boldsymbol{j} token \boldsymbol{j} layer 33 layer 34 layer 35 layer 36 15 15 0 0 0 0 100 100 100 -100 10 token 300 300 token 300 token 300 so token 300 a. 500 200 400400 400400 500 -500 $500 \cdot$ 500 200 400 200 400 200 400 200 400 0 0 token jtoken \boldsymbol{j} token \boldsymbol{j} token jlayer 37 layer 39 layer 40 layer 38 15 0 100 100 100 -100 - 10 - 10 token 300 token 300 300 token 300 300 oken 300 300 300 400 400400 400 500 -500 500 · 500 400 200 400200 400 200 200 400 0 0 token jtoken \boldsymbol{j} token \boldsymbol{j} token jlayer 41 layer 42 layer 43 layer 44 0 0 0 0 - 10 100 · 100 100 -100 token 300 300 oken 300 - 200 - 3 token 300 300 token 300 · • 200 · • 400 400400 400 500 -500 500 500 200 400200 400200 400 200 4000 0 token \boldsymbol{j} token \boldsymbol{j} token jtoken \boldsymbol{j} layer 45 layer 47 layer 46 layer 48 0 0 0 0 100 · 100 100 -100 token 300 300 token 300 300 token 300 - 200 oken 300 -400 400 400 400 500 -500 $500 \cdot$ 500 200 400 200 400 200 400 400 200

token j

token \boldsymbol{j}

token \boldsymbol{j}

token j