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 3 layer 1 layer 2 layer 4 token 300 token 300 token 100 token token s 300 -token jtoken jtoken jtoken  $\boldsymbol{j}$ layer 5 layer 6 layer 7 layer 8 100 -token 200 token so 200 oken 500 to 200 oken 200 token jtoken  $\boldsymbol{j}$ token jtoken  $\boldsymbol{j}$ layer 9layer 12 layer 10 layer 11 token 200 token 200 token 200 token 300 300 · token  $\boldsymbol{j}$ token  $\boldsymbol{j}$ token  $\boldsymbol{j}$ token  $\boldsymbol{j}$ layer 16 layer 13 layer 14 layer 15 token 200 token s. 200 token 200 token  $\boldsymbol{j}$ token  $\boldsymbol{j}$ token  $\boldsymbol{j}$ token jlayer 20 layer 17 layer 18 layer 19 token 200 token 200 -token jtoken  $\boldsymbol{j}$ token  $\boldsymbol{j}$ token  $\boldsymbol{j}$ layer 22 layer 21 layer 23 layer 24 0 -100 -100 -token 200 token i 200 token 500 चुं <u>5</u> 200 300 · 

token j

token j

token j

token j