Heatmaps of attention logits  $\frac{1}{\sqrt{d'}}\langle Qx_i, Kx_j\rangle$  $(\verb|imdb_albert-x|| arge-v2\_params-random\_dense-off_layers-48)$ layer 26 layer 27 layer 25 layer 28 token so 200 token 8. token 300 token 200 300 -token  $\boldsymbol{j}$ token  $\boldsymbol{j}$ token  $\boldsymbol{j}$ token  $\boldsymbol{j}$ layer 32 layer 29 layer 30 layer 31 100 -token 300 token i token i300 · token jtoken  $\boldsymbol{j}$ token  $\boldsymbol{j}$ token  $\boldsymbol{j}$ layer 33 layer 34 layer 35 layer 36 100 -token i token 300 token itoken 200 300 -token jtoken  $\boldsymbol{j}$ token  $\boldsymbol{j}$ token jlayer 37 layer 38 layer 39 layer 40 token 200 token 8. token 200 token i300 -token jtoken jtoken jtoken  $\boldsymbol{j}$ layer 41 layer 42 layer 43 layer 44 100 -token s token % token som 200 token itoken jtoken jtoken  $\boldsymbol{j}$ token  $\boldsymbol{j}$ layer 45 layer 46 layer 47 layer 48 100 -token i token i 200 token i 200 token 200 300 -token jtoken  $\boldsymbol{j}$ token  $\boldsymbol{j}$ token  $\boldsymbol{j}$