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 25 layer 26 layer 27 layer 28 token i. token 300 300 token itoken i400 -token  $\boldsymbol{j}$ token  $\boldsymbol{j}$ token  $\boldsymbol{j}$ token  $\boldsymbol{j}$ layer 29 layer 30 layer 31 layer 32 100 · to spot 300 and 300 token itoken itoken itoken  $\boldsymbol{j}$ token  $\boldsymbol{j}$ token jtoken  $\boldsymbol{j}$ layer 33 layer 34 layer 35 layer 36 100 -token itoken 300 token itoken i400 -token  $\boldsymbol{j}$ token jtoken  $\boldsymbol{j}$ token  $\boldsymbol{j}$ layer 39 layer 37 layer 40 layer 38 to de la serie de token itoken itoken i400 · token jtoken  $\boldsymbol{j}$ token  $\boldsymbol{j}$ token  $\boldsymbol{j}$ layer 41 layer 42 layer 43 layer 44 100 · 100 -token 300 : token itoken itoken i400 -token  $\boldsymbol{j}$ token jtoken jtoken jlayer 45 layer 46 layer 47 layer 48 100 · 100 -token 300 solution 300 token 300 token itoken i400 -token jtoken jtoken  $\boldsymbol{j}$ token  $\boldsymbol{j}$