Heatmaps of attention logits  $\frac{1}{\sqrt{d'}}\langle Qx_i, Kx_j\rangle$  $(\verb|imdb_a| bert-x| arge-v2\_params-random\_dense-off\_layers-48)$ layer 28 layer 25 layer 26 layer 27 token i 200 token itoken itoken itoken  $\boldsymbol{j}$ token  $\boldsymbol{j}$ token  $\boldsymbol{j}$ token  $\boldsymbol{j}$ layer 32 layer 29 layer 30 layer 31 100 · token 500 · 200 · token itoken itoken itoken  $\boldsymbol{j}$ token  $\boldsymbol{j}$ token  $\boldsymbol{j}$ token  $\boldsymbol{j}$ layer 33 layer 34 layer 35 layer 36 100 · token *i* 500 token itoken itoken itoken jtoken  $\boldsymbol{j}$ token jtoken  $\boldsymbol{j}$ layer 39 layer 37 layer 40 layer 38 100 -token i 200 token itoken itoken itoken jtoken jtoken  $\boldsymbol{j}$ token  $\boldsymbol{j}$ layer 41 layer 42 layer 44 layer 43 100 -100 -token 8. token itoken 5. token i400 -token jtoken jtoken jtoken  $\boldsymbol{j}$ layer 45 layer 46 layer 47 layer 48 100 -100 -token i 200 token *i*. token *i*. token i300 -token  $\boldsymbol{j}$ token  $\boldsymbol{j}$ token  $\boldsymbol{j}$ token  $\boldsymbol{j}$