

Heatmaps of attention logits  $\frac{1}{\sqrt{d}} \langle Qx_i, Kx_j \rangle$   
(uniform\_bert-large-uncased\_params-trained\_dense-off\_layers-24, sample 8)

