Heatmaps of attention logits $\frac{1}{\sqrt{d'}}\langle Qx_i, Kx_j\rangle$ $({\tt randomtext_albert_xlarge_v2_params_trained_dense_off_layers_48})$ layer 25 layer 26 layer 27 layer 28 0 0 100 100 -100 100 - 10 10 token 300 300 oken 300 300 oken 300 300 token 300 200 400 400 400400500 -500 -500500 200400 200 400200400 200400token jtoken \boldsymbol{j} token \boldsymbol{j} token \boldsymbol{j} layer 29 layer 30 layer 31 layer 32 0 0 100 100 100 100 10 roken 300 300 token 300 300 token 300 300 token 300 300 400 400400 400 500 -500 -500 500 -200 400 200 400 200 400 200 0 400token \boldsymbol{j} token jtoken jtoken \boldsymbol{j} layer 34 layer 33 layer 35 layer 36 0 0 10 10 100 100 -100 100 token 300 300 oken 300 300 - 5 to ken so 300 a 200 oken 300 300 400400400400 500 500 -500 500 400 400 200 200 400200 200 4000 0 token jtoken \boldsymbol{j} token \boldsymbol{j} token jlayer 37 layer 38 layer 39 layer 40 0 0 0 100 100 100100 10 token 300 so 200 token 300 ≈ 300 token 300 200 token 300 a. 300 400 -400400400 500 -500 -500 500 400 200 400200 400200 200 4000 token jtoken jtoken \boldsymbol{j} token \boldsymbol{j} layer 41 layer 42 layer 43 layer 44 0 0 100 · 100 -100 100 - 10 **-** 10 token 300 300 token 300 300 token 300 · · · 200 · · oken 300 300 400 -400 400 400 500 -500 -500 500 -200 400200 400 200 400200 400 0 0 0 0 token jtoken \boldsymbol{j} token \boldsymbol{j} token \boldsymbol{j} layer 45 layer 46 layer 47 layer 48 0 0 0 0 100 -100 -100 100 10 10 - 10 → 200 300 token 300 -± 200 september token 300 a. 300 400 400400400 500 -500 -500 200400200400200 400 200 4000

token j

token \boldsymbol{j}

token j

token \boldsymbol{j}