Decrease in token probability after ablating attention head 0 0 0.08 Decrease in token probability 0.06 0.04 0.02 0.00 -0.02Attn Head Activettn Head Inactive Overall