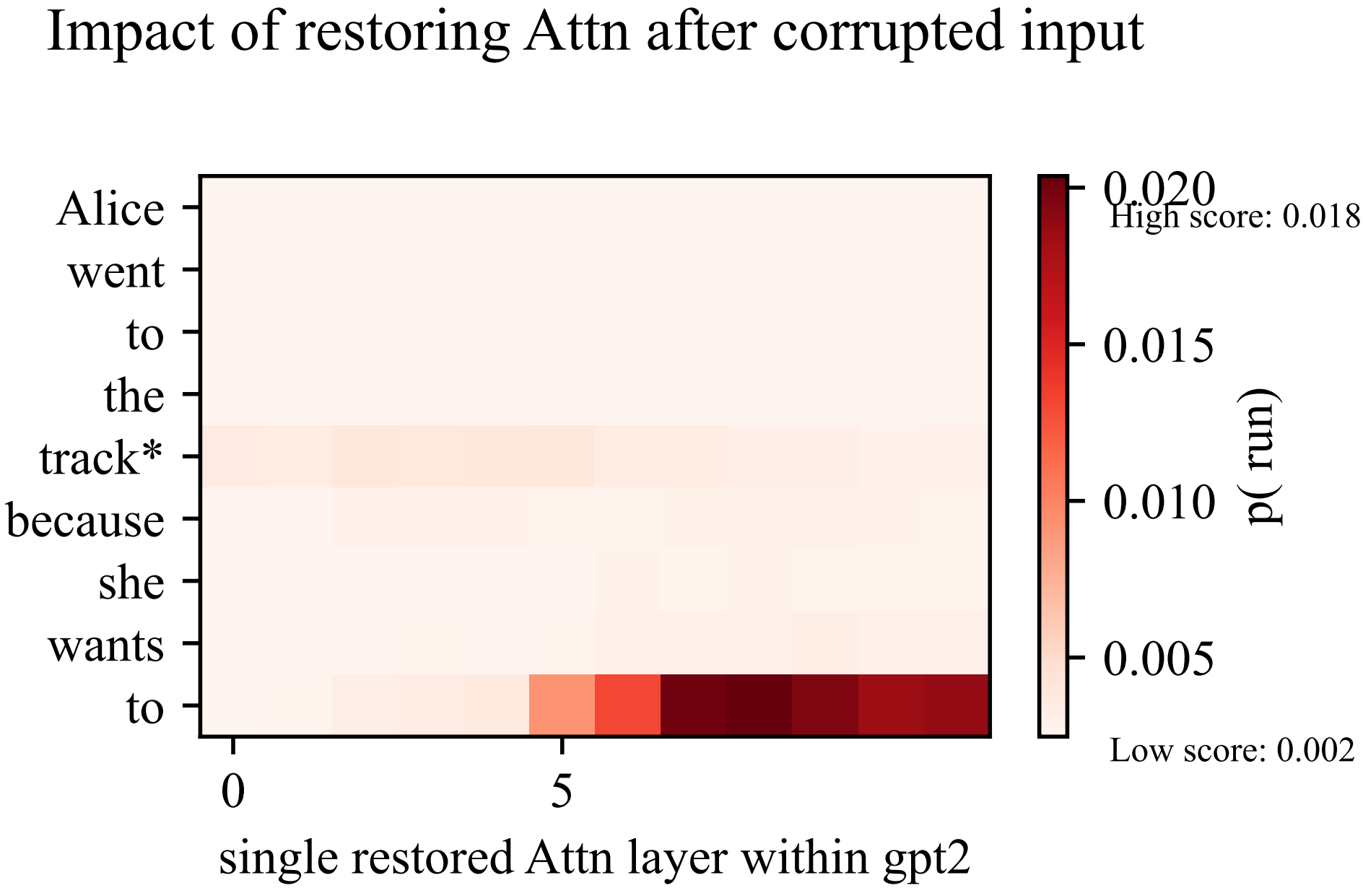
Impact of restoring Attn after corrupted input
Alice
went
to
the
track*
because
she
wants
to
0
5
single restored Attn layer within gpt2
0.020
High score: 0.018
0.015
0.010
0.005
Low score: 0.002
p( run)