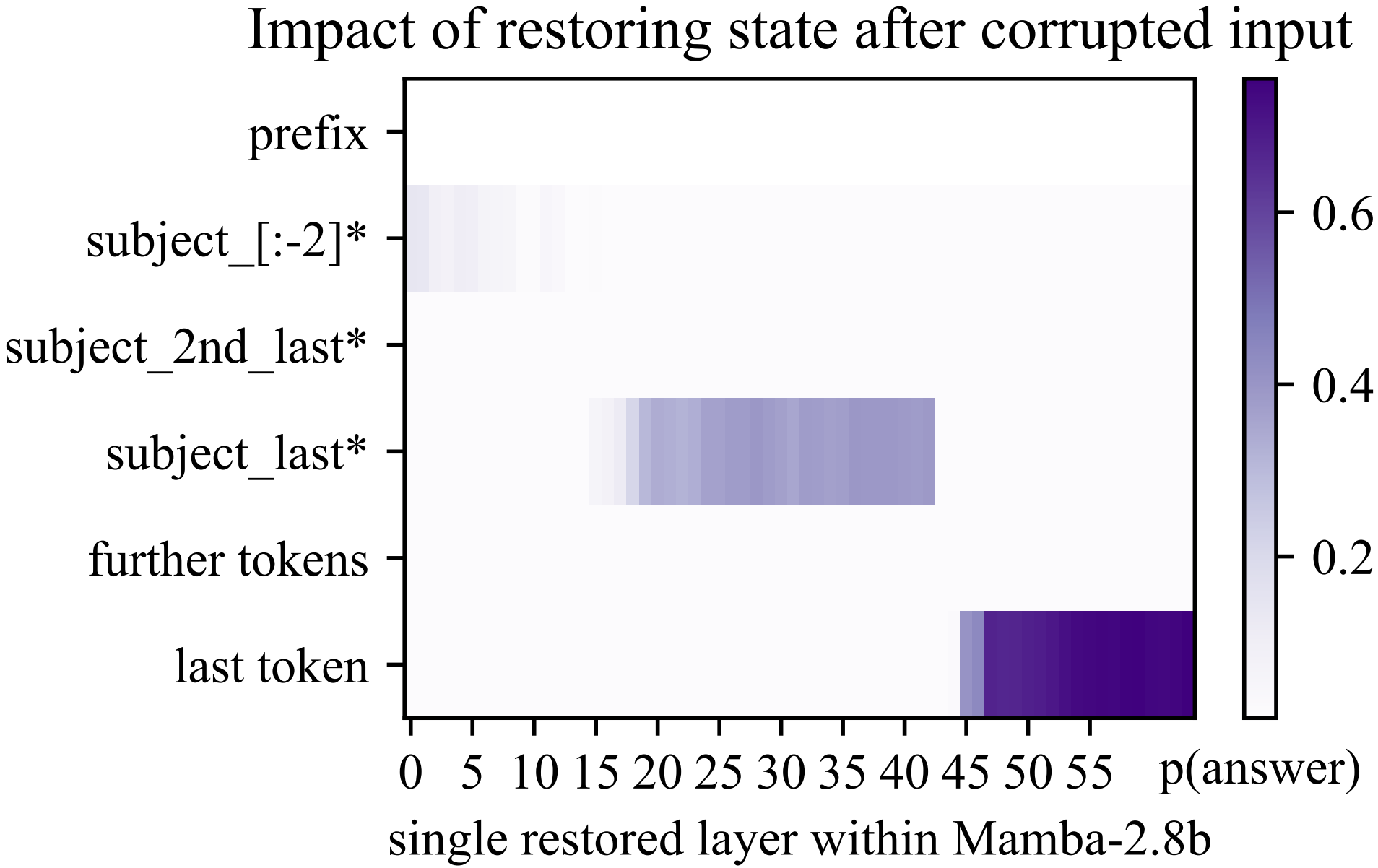
Impact of restoring state after corrupted input
prefix
subject_[:-2]*
subject_2nd_last*
subject_last*
further tokens
last token
0
5
10
15
20
25
30
35
40
45
50
55
single restored layer within Mamba-2.8b
0.6
0.4
0.2
p(answer)