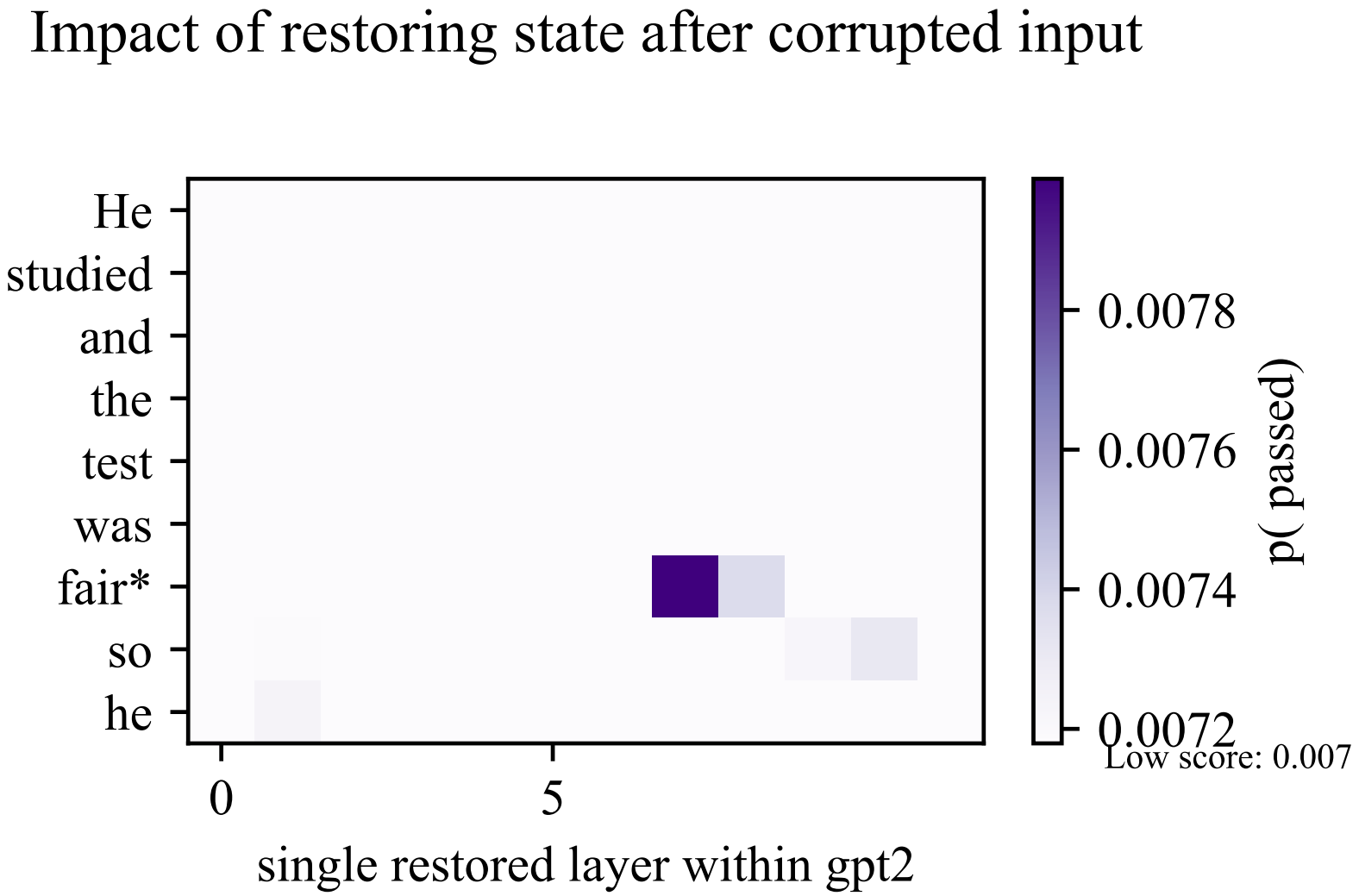
Impact of restoring state after corrupted input
He
studied
and
the
test
was
fair*
so
he
0
5
single restored layer within gpt2
0.0078
0.0076
0.0074
0.0072
p(passed)
Low score: 0.007
High score: 0.006