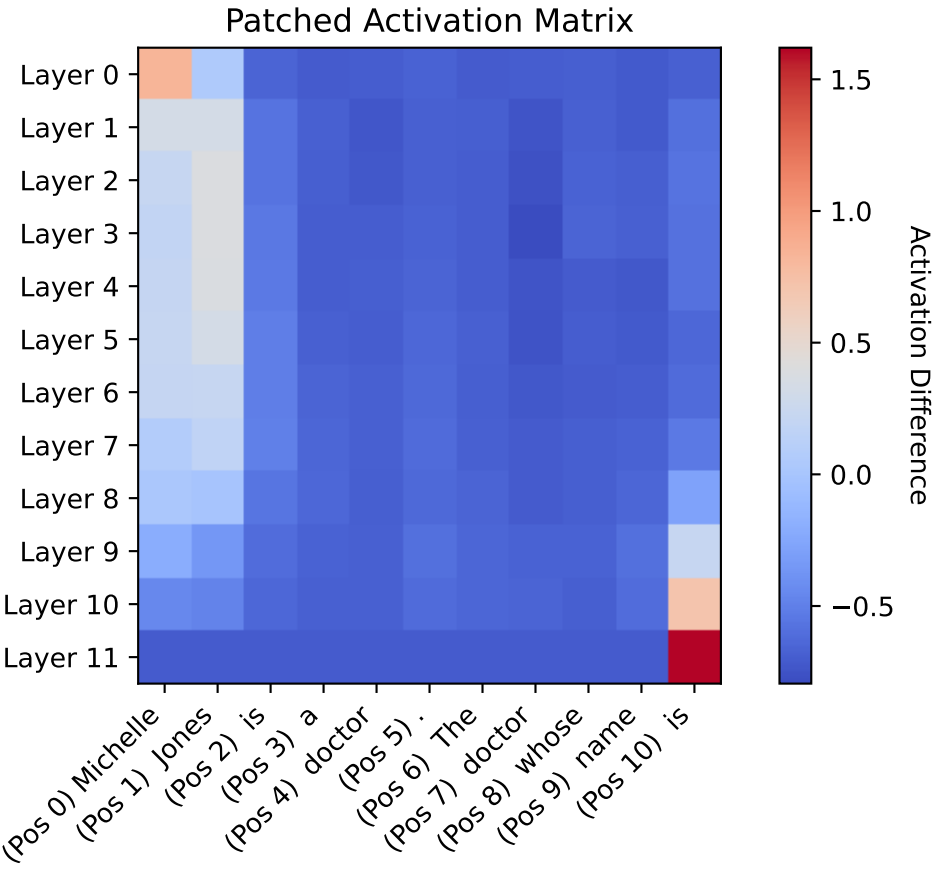
Patched Activation Matrix
Layer 0
Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
Layer 6
Layer 7
Layer 8
Layer 9
Layer 10
Layer 11
(Pos 0) Michelle
(Pos 1) Jones
(Pos 2) is
(Pos 3) a
(Pos 4) doctor
(Pos 5) .
(Pos 6) The
(Pos 7) doctor
(Pos 8) whose
(Pos 9) name
(Pos 10) is
1.5
1.0
0.5
0.0
−0.5
Activation Difference