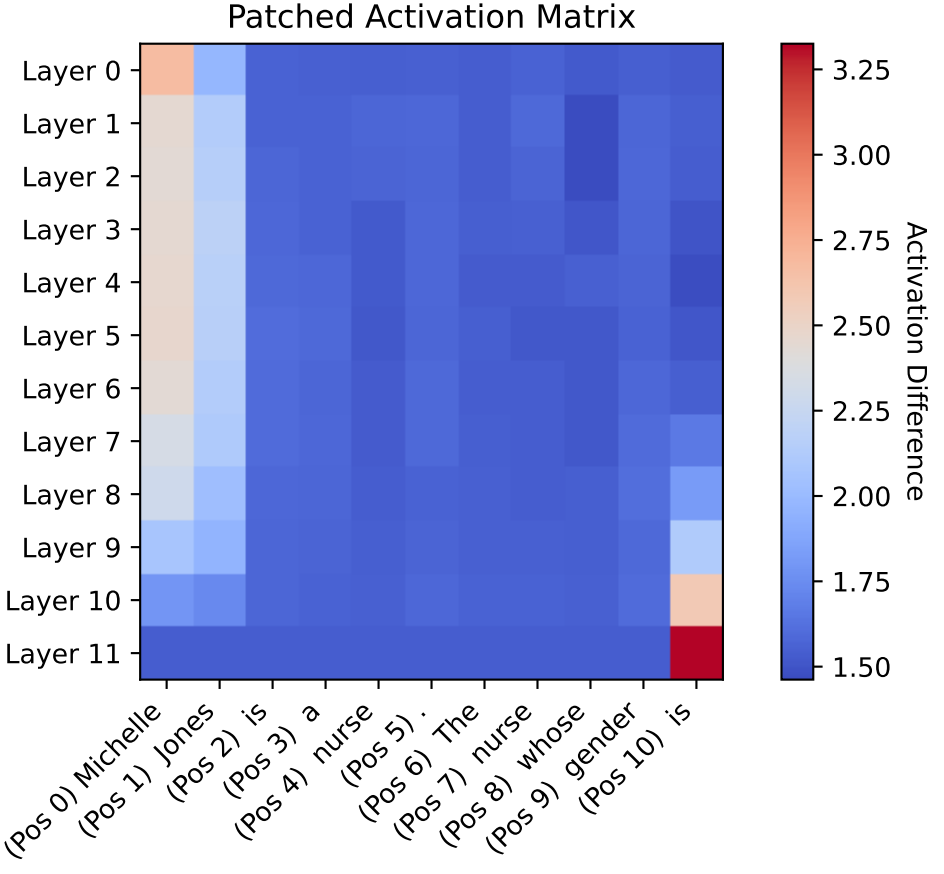
Patched Activation Matrix
Layer 0
Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
Layer 6
Layer 7
Layer 8
Layer 9
Layer 10
Layer 11
(Pos 0) Michelle
(Pos 1) Jones
(Pos 2) is
(Pos 3) a
(Pos 4) nurse
(Pos 5) .
(Pos 6) The
(Pos 7) nurse
(Pos 8) whose
(Pos 9) gender
(Pos 10) is
3.25
3.00
2.75
2.50
2.25
2.00
1.75
1.50
Activation Difference