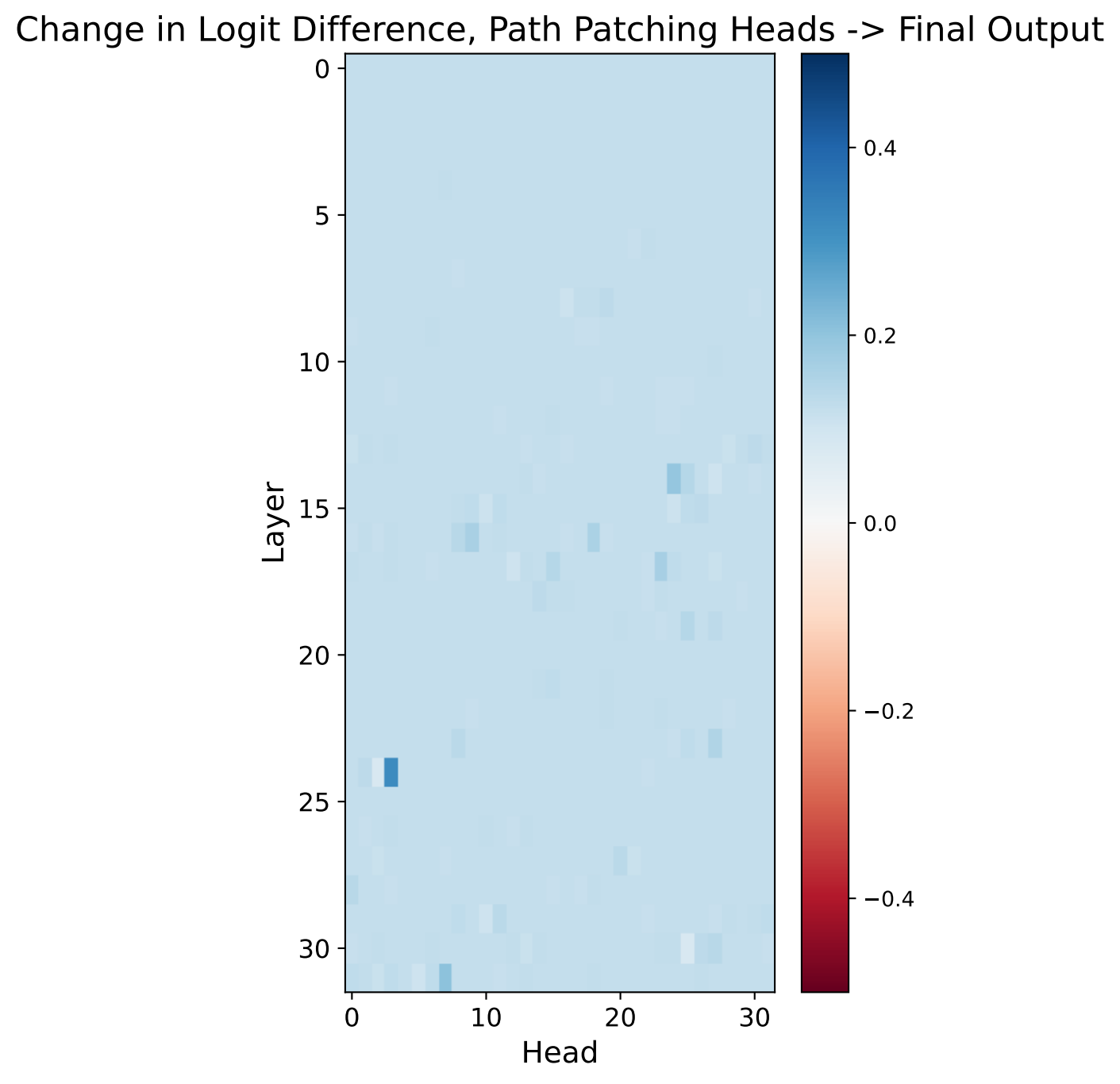
Change in Logit Difference, Path Patching Heads -> Final Output
Layer
Head
0
5
10
15
20
25
30
0
10
20
30
0.4
0.2
0.0
−0.2
−0.4