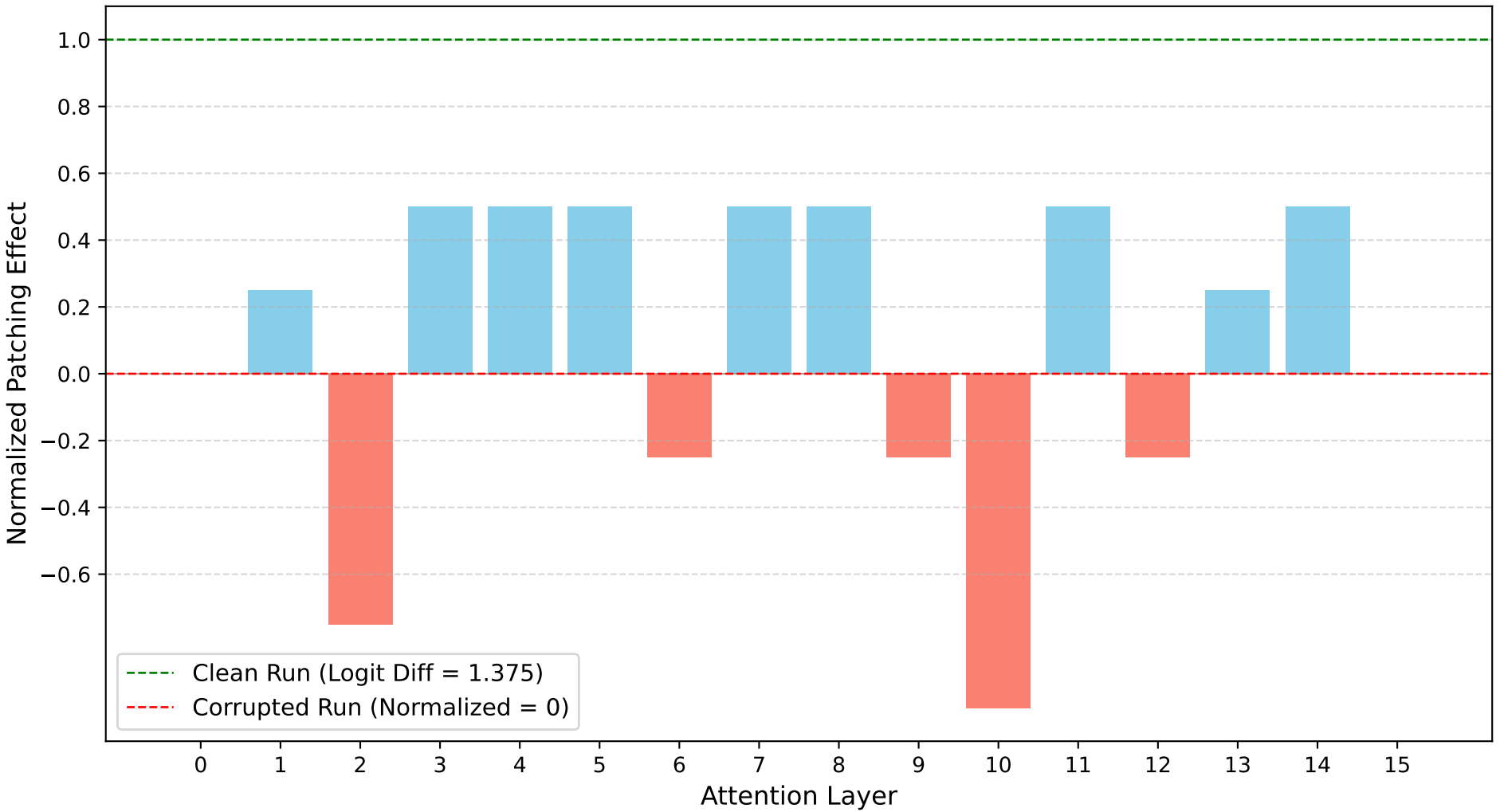
Effect of Patching Attention Layers on Logit Difference
(Clean: "Paris" vs Corrupted: "Berlin")
Normalized Patching Effect
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
Clean Run (Logit Diff = 1.375)
Corrupted Run (Normalized = 0)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Attention Layer