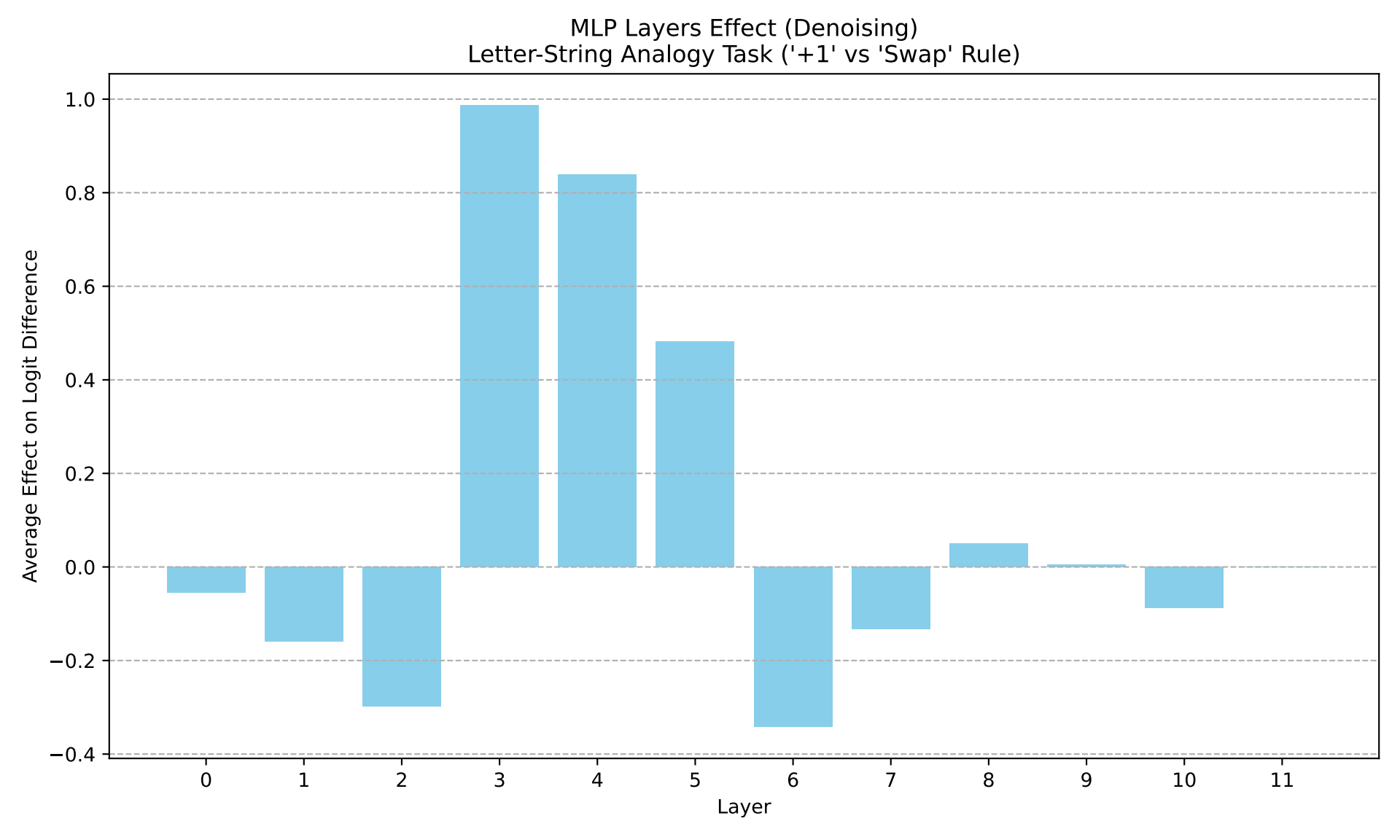
MLP Layers Effect (Denoising)
Letter-String Analogy Task ('+1' vs 'Swap' Rule)
Average Effect on Logit Difference
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
1
2
3
4
5
6
7
8
9
10
11
Layer