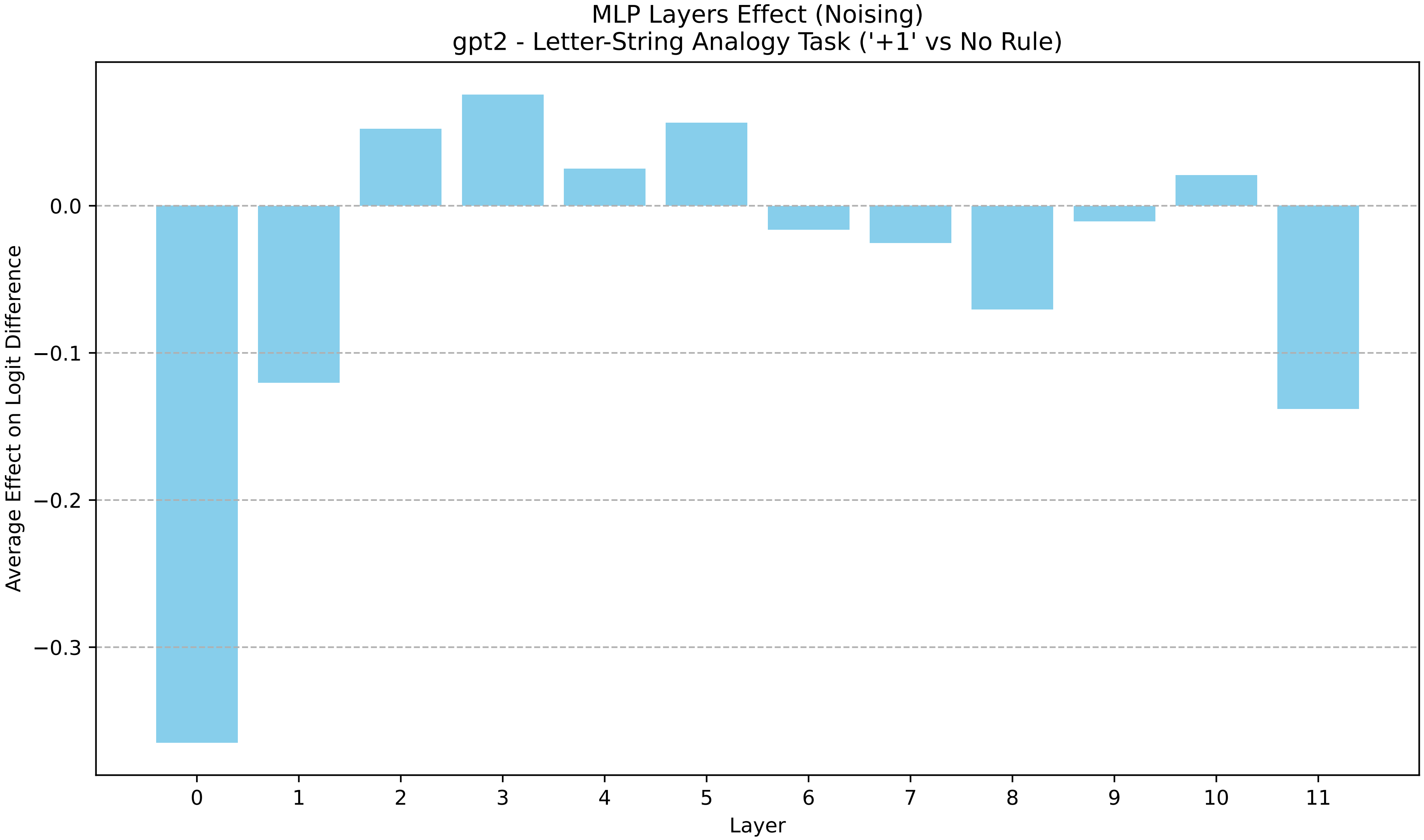
MLP Layers Effect (Noising)
gpt2 - Letter-String Analogy Task ('+1' vs No Rule)
Average Effect on Logit Difference
0.0
−0.1
−0.2
−0.3
0
1
2
3
4
5
6
7
8
9
10
11
Layer