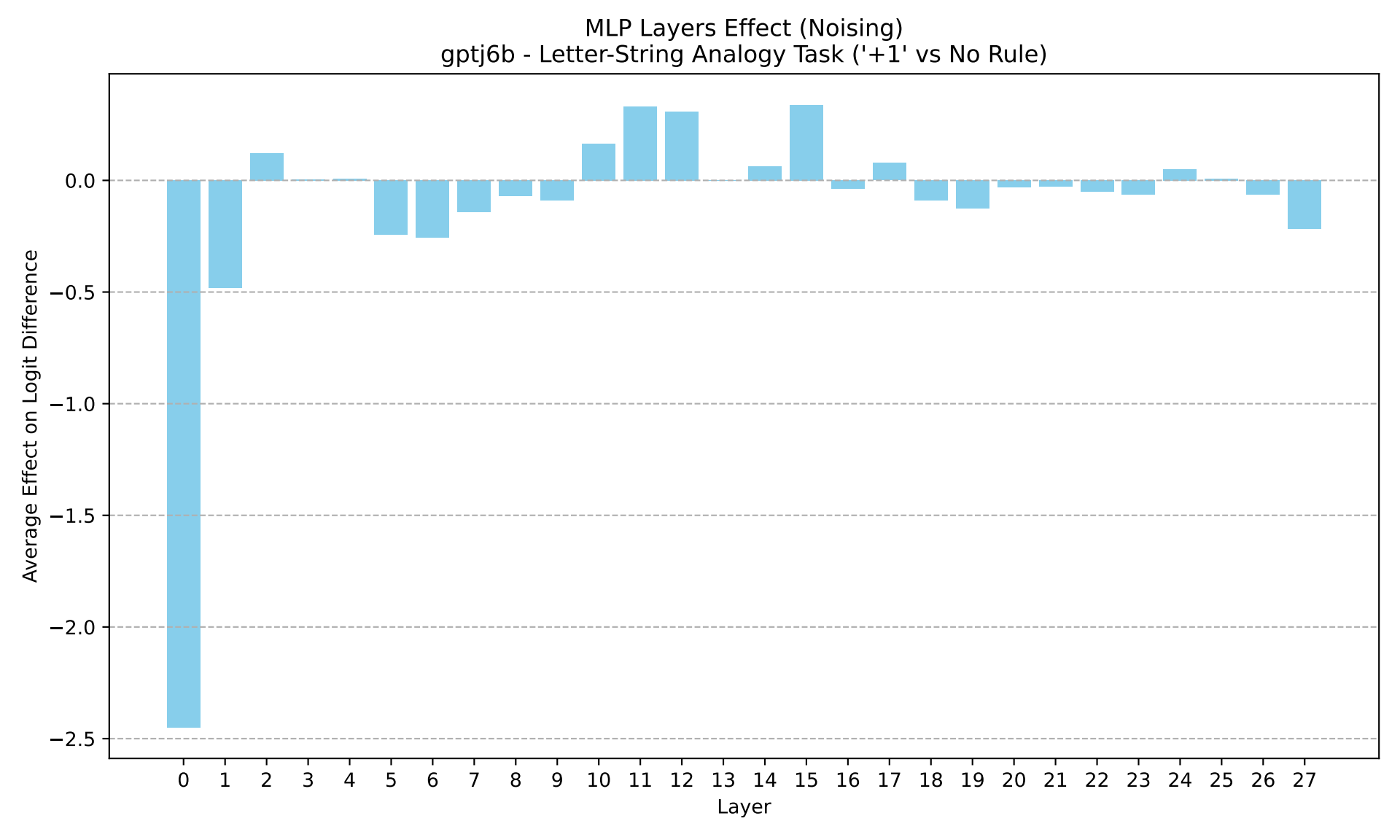
MLP Layers Effect (Noising)
gptj6b - Letter-String Analogy Task ('+1' vs No Rule)
Average Effect on Logit Difference
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
Layer