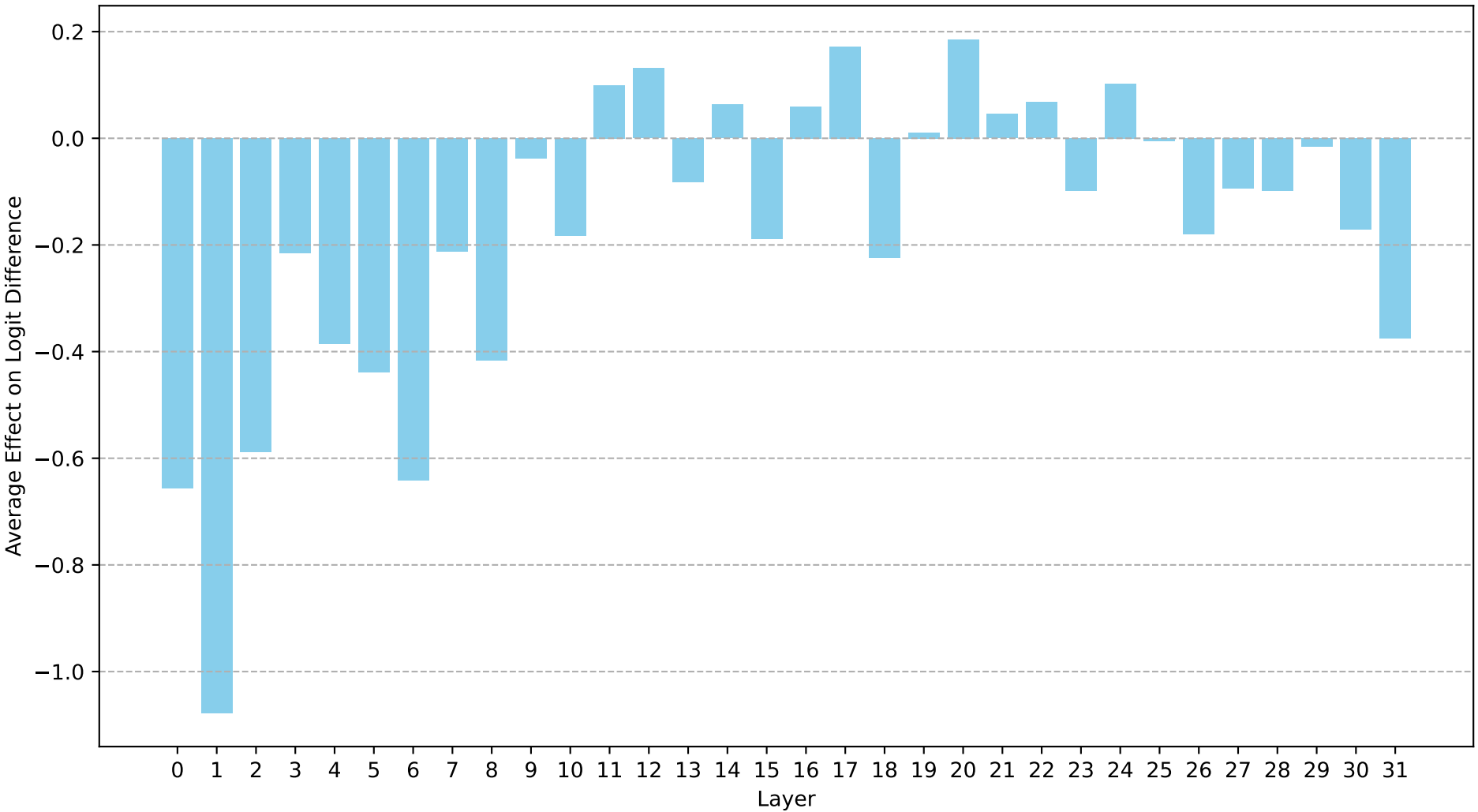
MLP Layers Effect (Noising)
llama27b - Letter-String Analogy Task ('+1' vs No Rule)
Average Effect on Logit Difference
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
Layer