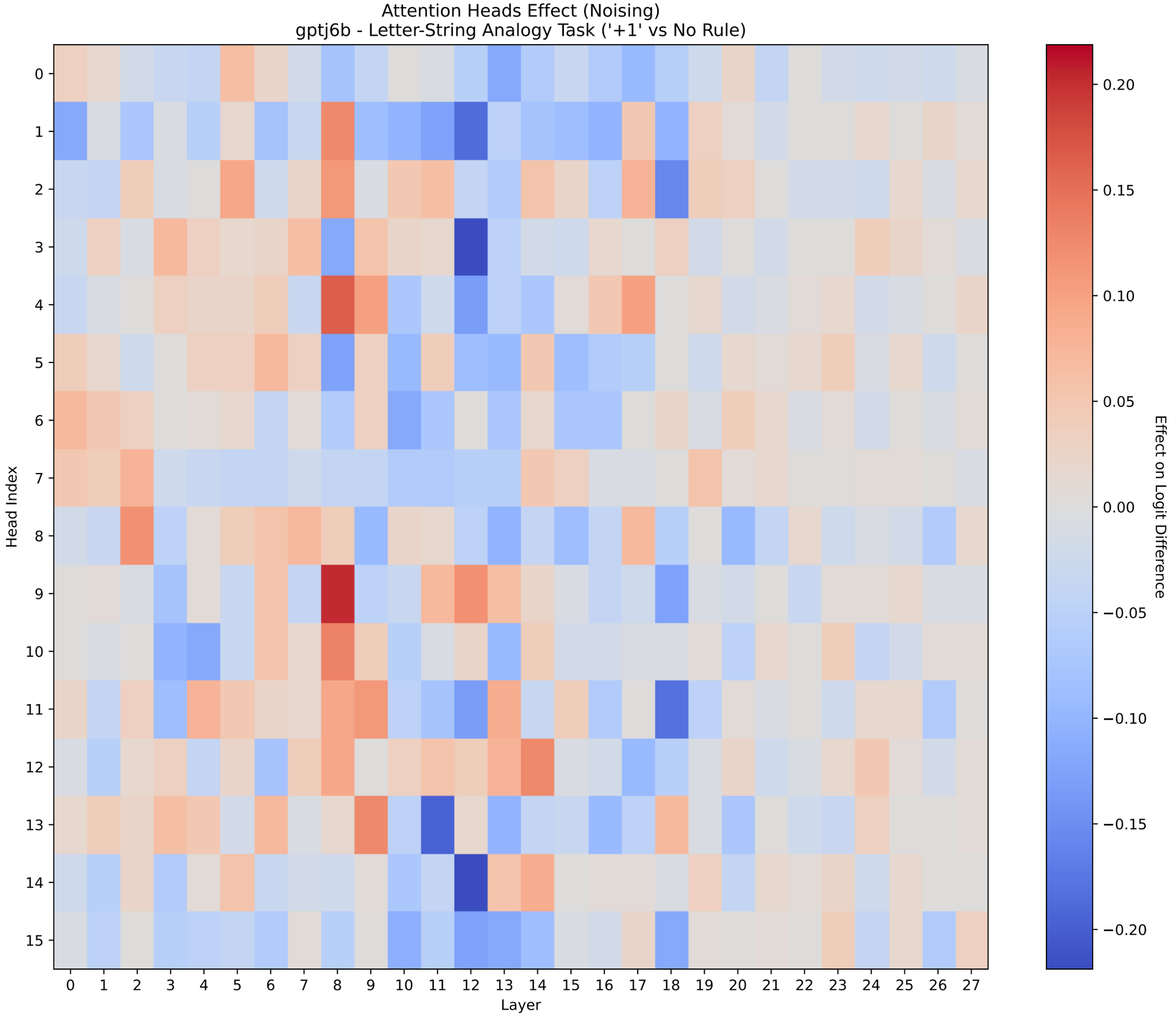
Attention Heads Effect (Noising)
gptj6b - Letter-String Analogy Task ('+1' vs No Rule)
Head Index
Layer
Effect on Logit Difference
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20