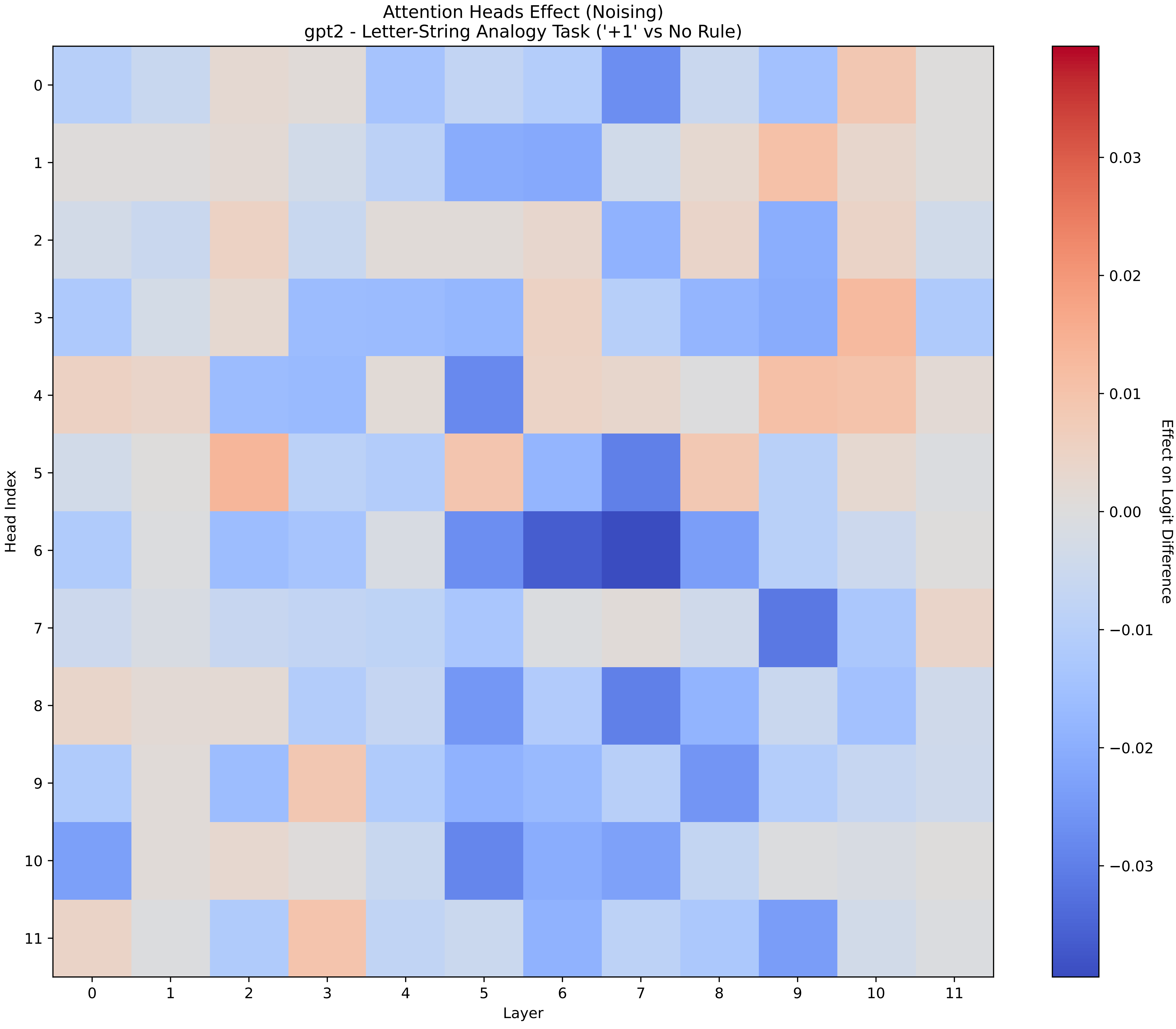
Attention Heads Effect (Noising)
gpt2 - Letter-String Analogy Task ('+1' vs No Rule)
Head Index
0
1
2
3
4
5
6
7
8
9
10
11
Layer
0
1
2
3
4
5
6
7
8
9
10
11
Effect on Logit Difference
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03