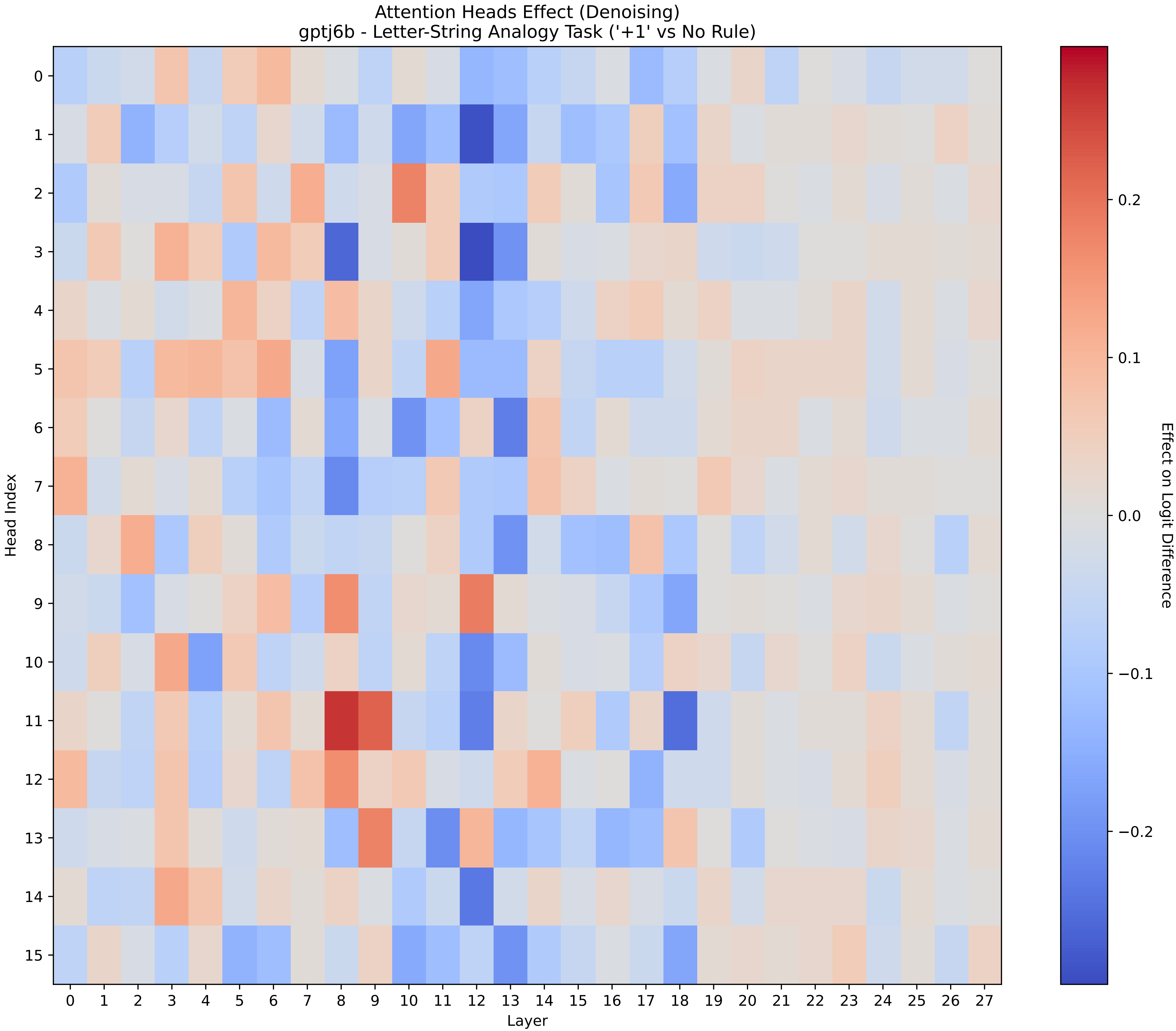
Attention Heads Effect (Denoising)
gptj6b - Letter-String Analogy Task ('+1' vs No Rule)
Head Index
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Layer
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
Effect on Logit Difference
0.2
0.1
0.0
−0.1
−0.2