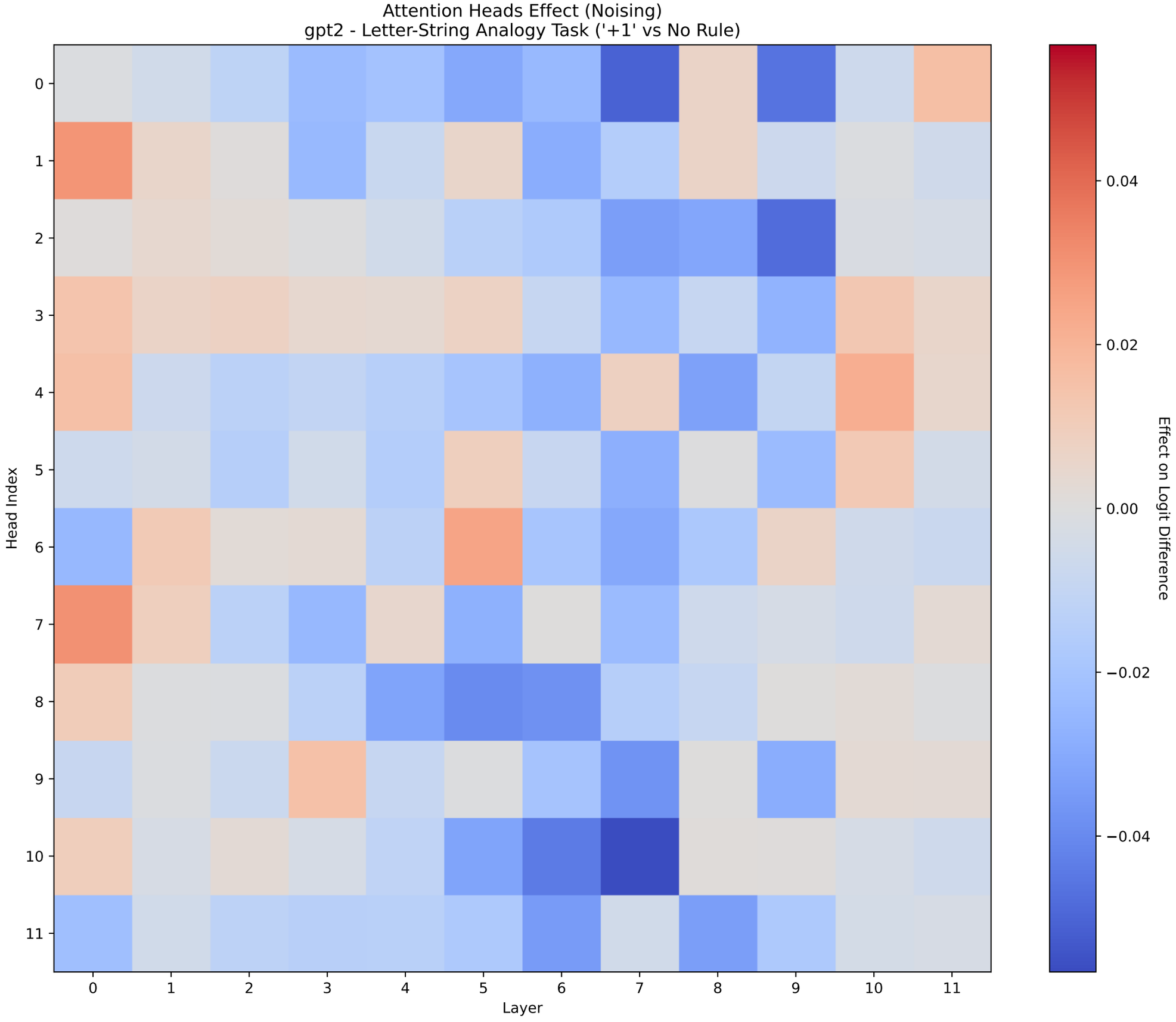
Attention Heads Effect (Noising)
gpt2 - Letter-String Analogy Task ('+1' vs No Rule)
Head Index
Layer
Effect on Logit Difference
0.04
0.02
0.00
−0.02
−0.04