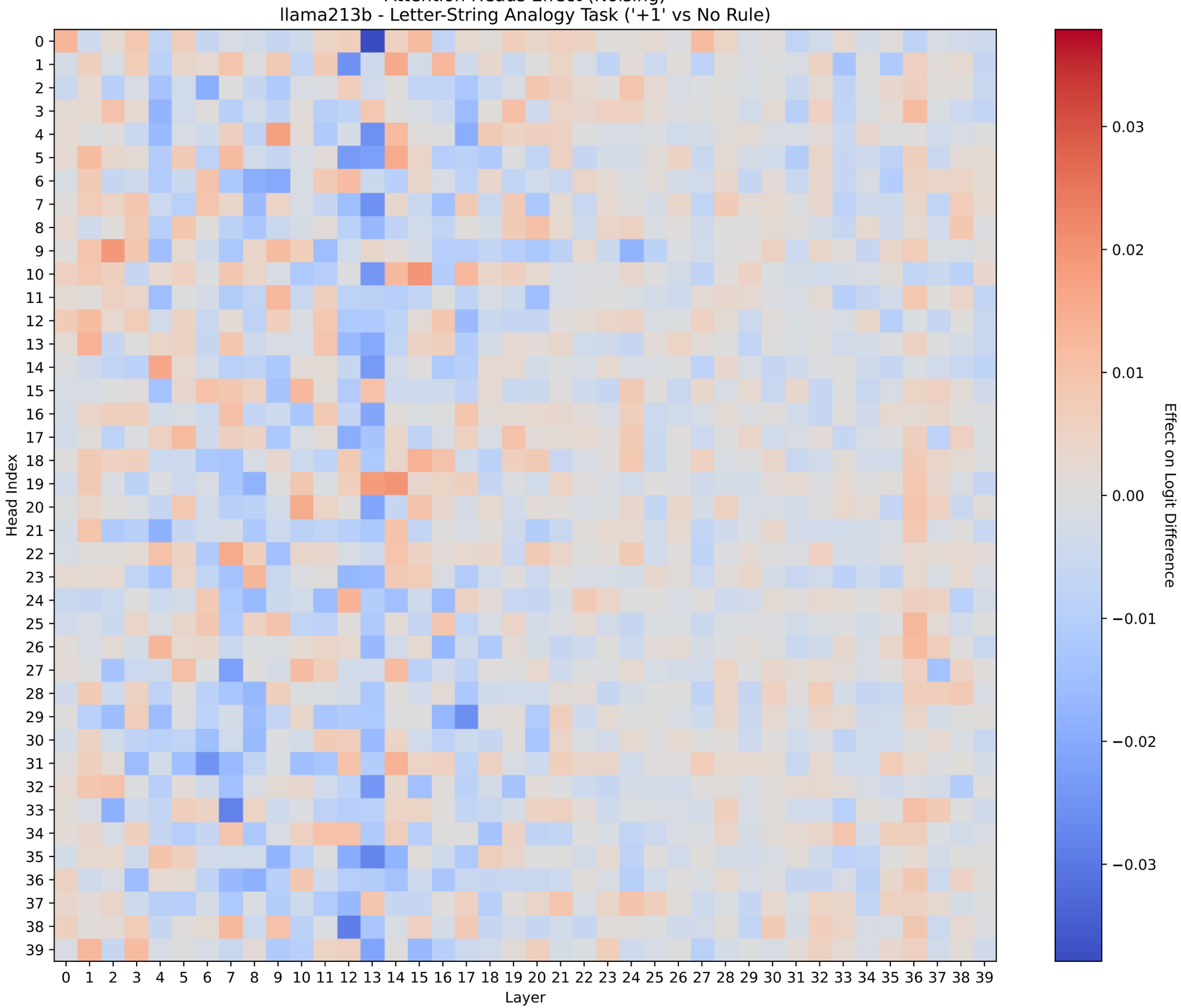
Attention Heads Effect (Noising)
llama213b - Letter-String Analogy Task ('+1' vs No Rule)
Head Index
Layer
Effect on Logit Difference
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03