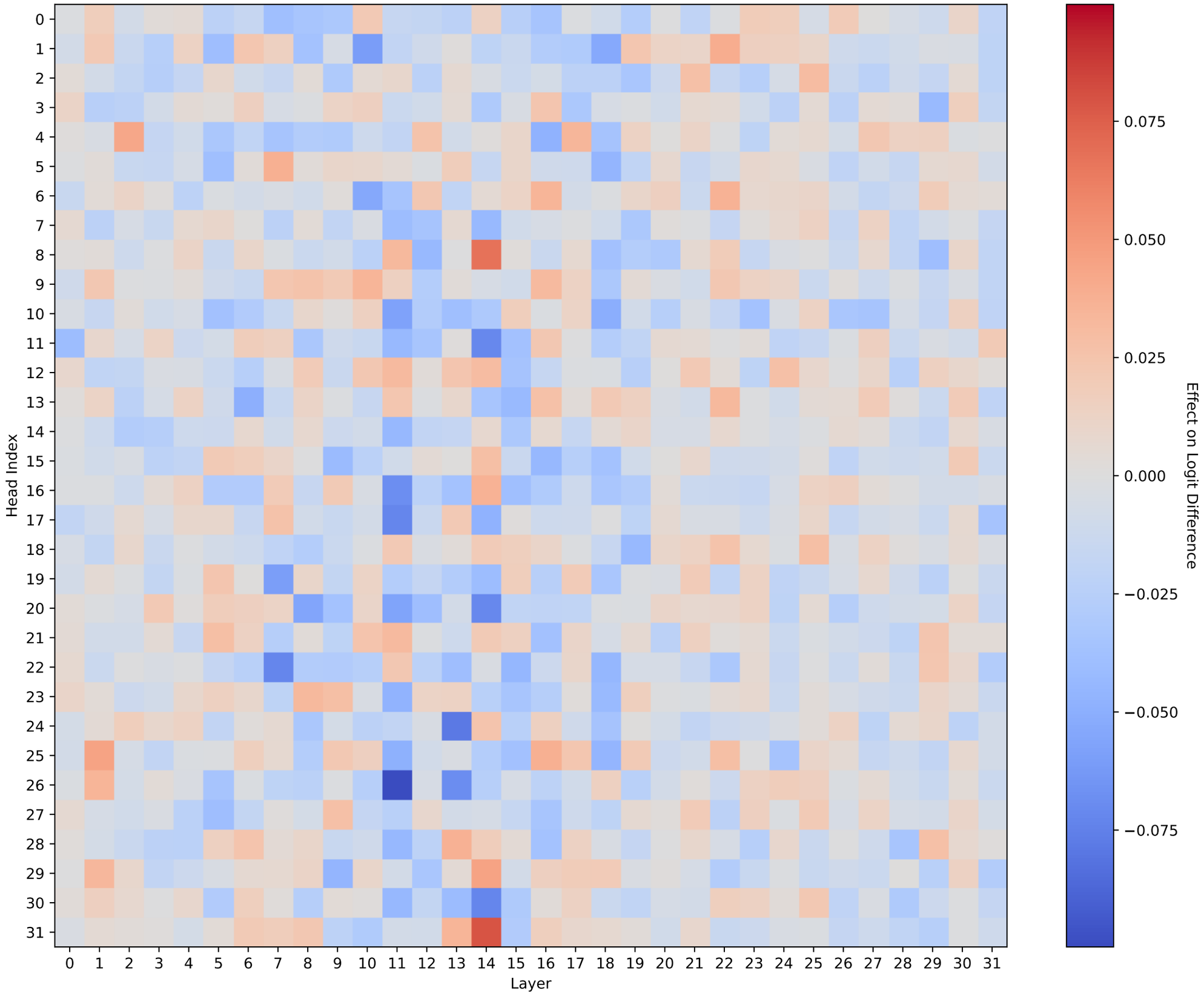
Attention Heads Effect (Noising)
llama27b - Letter-String Analogy Task ('+1' vs No Rule)
Head Index
Layer
Effect on Logit Difference
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075