Attention Heads Effect (Noising)
llama27b - Letter-String Analogy Task ('+1' vs No Rule) 0 1 - 0.03 3 5 6 - 0.02 7 8 9 10 0.01 11 -12 Effect on Logit Difference 13 -14 - 29 15 - 29 16 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 29 17 - 2 - 0.00 18 19 20 -0.0121 22 -23 -24 - -0.02 25 26 27 28 29 - 0.03 30 31 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 8 Layer