Attention Heads Effect (Noising)
llama27b - Letter-String Analogy Task ('+1' vs No Rule) 0 1 3 -- 0.075 6 7 0.050 8 9 10 11 - 0.025 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 - 0.000 18 19 -0.025 20 21 22 -23 --0.050 24 25 26 27 -0.075 28 29 -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 Layer