Attention Heads Effect (Noising)
llama27b - Letter-String Analogy Task ('+1' vs No Rule) 0.04 0 1 3 0.03 5 6 7 0.02 8 9 10 11 - 0.01 12 Effect on Logit Difference 13 -Head Ide Head Ide 16 17 - 0.00 18 19 -0.0120 21 22 -23 -- 0.02 24 25 26 -27 - -0.03 28 29 -30 31 -0.049 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 2 8 Layer