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