Attention Heads Effect (Denoising)
llama27b - Letter-String Analogy Task ('+1' vs No Rule) - 0.100 0 1 -0.075 0.050 10 -11 -- 0.025 12 -13 -14 -Head Index Head Index 16 -17 -- 0.000 18 19 --0.025 20 -21 -22 -23 -- -0.050 24 -25 · 26 -27 --0.075 28 -29 -30 31 L -0.100 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 Layer