Attention Heads Effect (Noising)
Ilama213b - Letter-String Analogy Task ('+1' vs No Rule) 0.03 0.02 - 0.01 Effect on Logit Difference Head Index 19 20 21 - 0.00 -0.01 - 0.02 - 0.03 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39

Layer