Attention Heads Effect (Noising)
gptj6b - Letter-String Analogy Task ('+1' vs No Rule) 0 · 0.20 1 · 0.15 2 3 - 0.10 4 5 - 0.05 6 Effect on Logit Difference Head Index 0.00 8 9 -0.05 10 11 -0.10 12 -13 -0.1514 -0.20 15 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 9 Layer