Attention Heads Effect (Noising)
gptj6b - Letter-String Analogy Task ('+1' vs No Rule) 0 · 1 - 0.06 2 3 0.04 4 5 · 0.02 6 -Effect on Logit Difference Head Index 0.00 8 9 -0.0210 11 -0.04 12 · 13 --0.06 14 15 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 2 3 5 8 9 Layer