Attention Heads Effect (Noising)
gpt2 - Letter-String Analogy Task ('+1' vs No Rule) 0 -- 0.03 1 -2 -- 0.02 3 -- 0.01 4 -Effect on Logit Difference 5 · Head Index 0.00 6 7 --0.01 8 -- -0.02 9 10 -- 0.03 11 -2 3 6 7 9 10 11 0 1 4 5 8 Layer