Attention Heads Effect (Noising)
gpt2 - Letter-String Analogy Task ('+1' vs No Rule) 0.04 0 1 -- 0.03 2 -0.02 3 -4 · - 0.01 Effect on Logit Difference 5 Head Index 0.00 6 -0.01 7 -8 --0.02 9 - -0.03 10 -11 --0.041 2 3 6 7 9 10 11 0 4 5 8 Layer