Attention Heads Effect (Noising)
gpt2 - Letter-String Analogy Task ('+1' vs No Rule) 0 0.06 1 -2 -- 0.04 3 -- 0.02 4 -Effect on Logit Difference 5 Head Index 0.00 6 7 --0.02 8 --0.04 9 10 -- 0.06 11 -3 7 9 10 11 0 1 2 4 5 6 8 Layer