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