Attention Heads Effect (Noising)
Letter-String Analogy Task ('+1' vs 'Swap' Rule) - 0.3 0 -1 -- 0.2 2 -3 -- 0.1 4 -Effect on Logit Difference 5 -Layer 0.0 6 7 --0.1 8 -9 -0.2 10 -11 --0.3Ó 3 7 9 10 11 1 2 4 5 6 8 Head Index