Attention Heads Effect (Denoising)
Letter-String Analogy Task ('+1' vs 'Swap' Rule) 0 - 0.15 1 -2 -- 0.10 3 -- 0.05 4 · Effect on Logit Difference 5 · 0.00 6 7 --0.05 8 -- 0.10 9 10 -- -0.15 11 -Ó i 2 3 4 7 9 10 11 5 6 8 Head Index