Attention Heads Effect (Denoising)
Letter-String Analogy Task ('+1' vs No Rule) 0 - 0.10 1 · 2 -3 -- 0.05 4 -Effect on Logit Difference 5 Head Index 0.00 6 7 --0.05 8 -9 10 -- -0.10

7

8

6

9

10

11

11 -

Ó

1

2

3

4

5

Layer