

Loss Difference From Intervention

0.010
0.005
0.000
-0.005
-0.010
-0.015

0.5

0.6

0.7

0.8

0.9

Mean Attention Weights on Extreme Tokens



wikipedia



github

