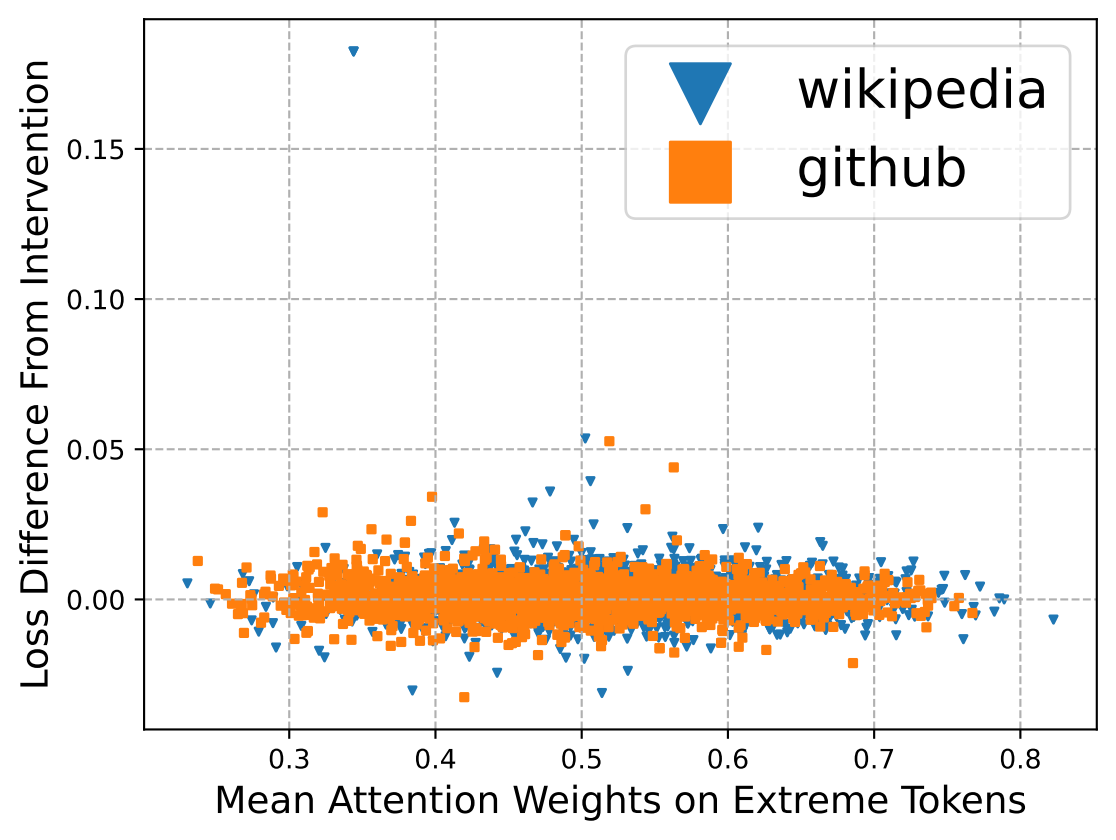
wikipedia
github
Loss Difference From Intervention
0.15
0.10
0.05
0.00
0.3
0.4
0.5
0.6
0.7
0.8
Mean Attention Weights on Extreme Tokens