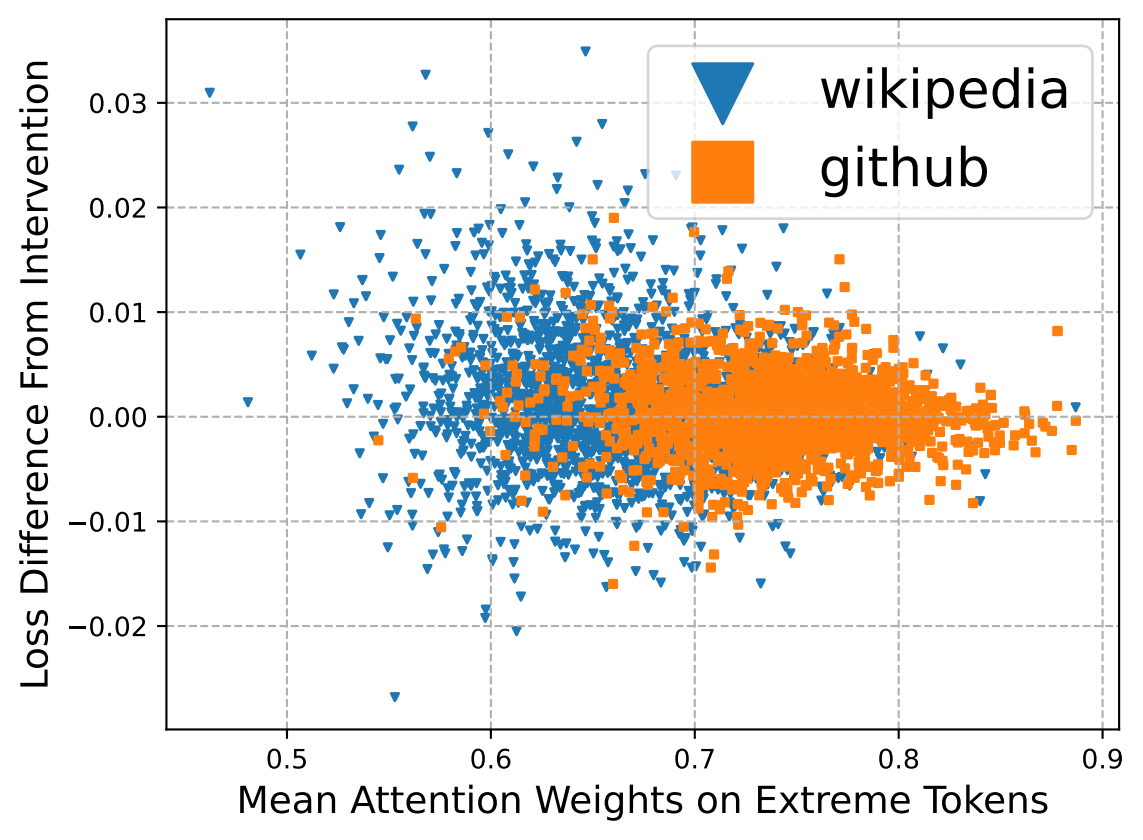
wikipedia
github
Loss Difference From Intervention
0.03
0.02
0.01
0.00
−0.01
−0.02
0.5
0.6
0.7
0.8
0.9
Mean Attention Weights on Extreme Tokens