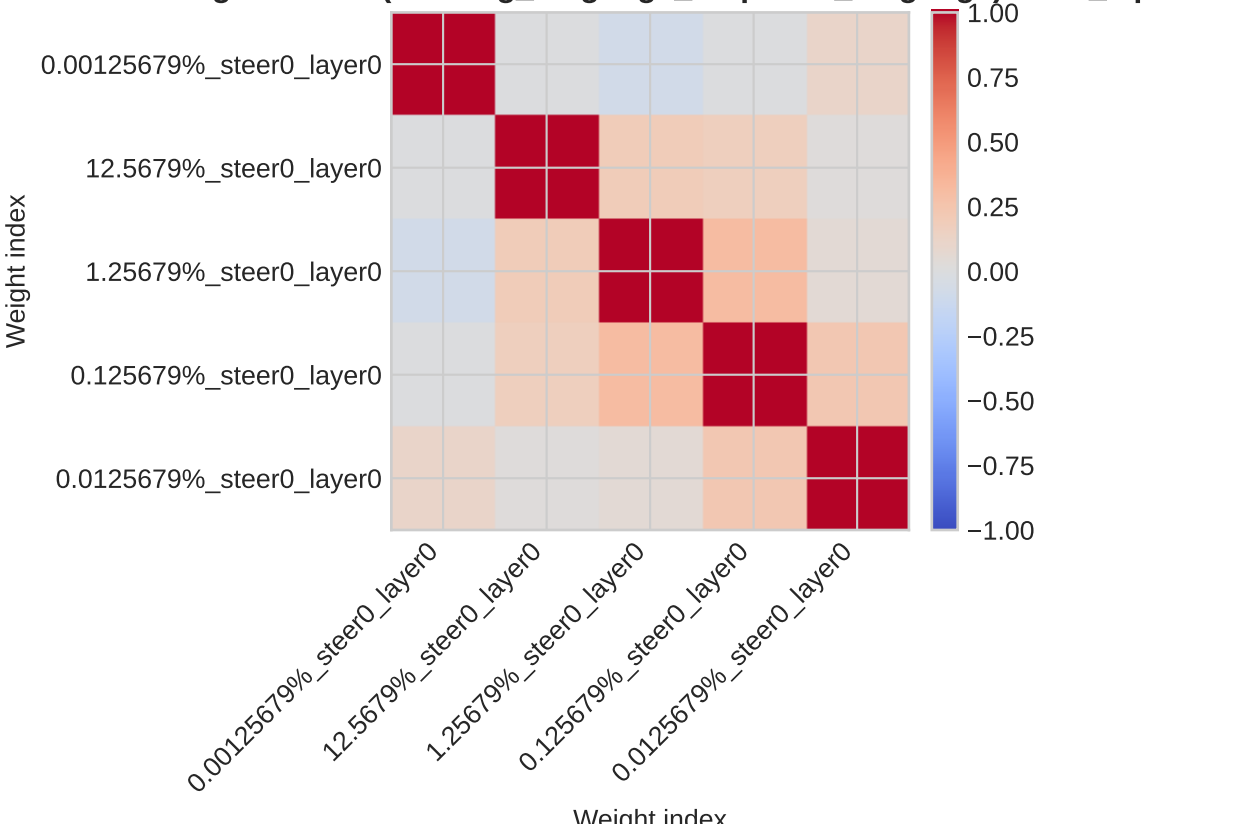
Probe weight cosine (steering_language_response_language) steer_0 probe0
0.00125679%_steer0_layer0
12.5679%_steer0_layer0
1.25679%_steer0_layer0
0.125679%_steer0_layer0
0.0125679%_steer0_layer0
Weight index
Weight index
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00