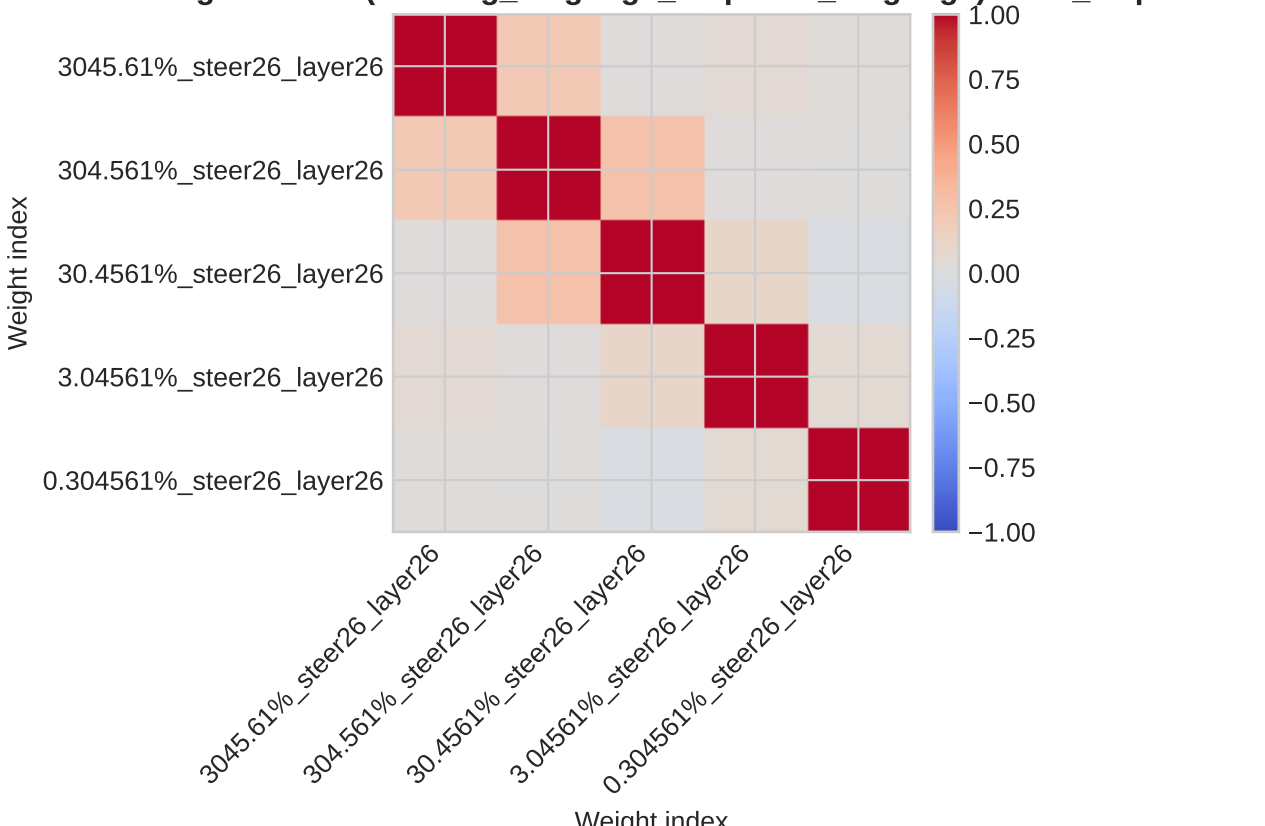
Probe weight cosine (steering_language_response_language) steer_26 probe26
3045.61%_steer26_layer26
304.561%_steer26_layer26
30.4561%_steer26_layer26
3.04561%_steer26_layer26
0.304561%_steer26_layer26
Weight index
3045.61%_steer26_layer26
304.561%_steer26_layer26
30.4561%_steer26_layer26
3.04561%_steer26_layer26
0.304561%_steer26_layer26
Weight index
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00