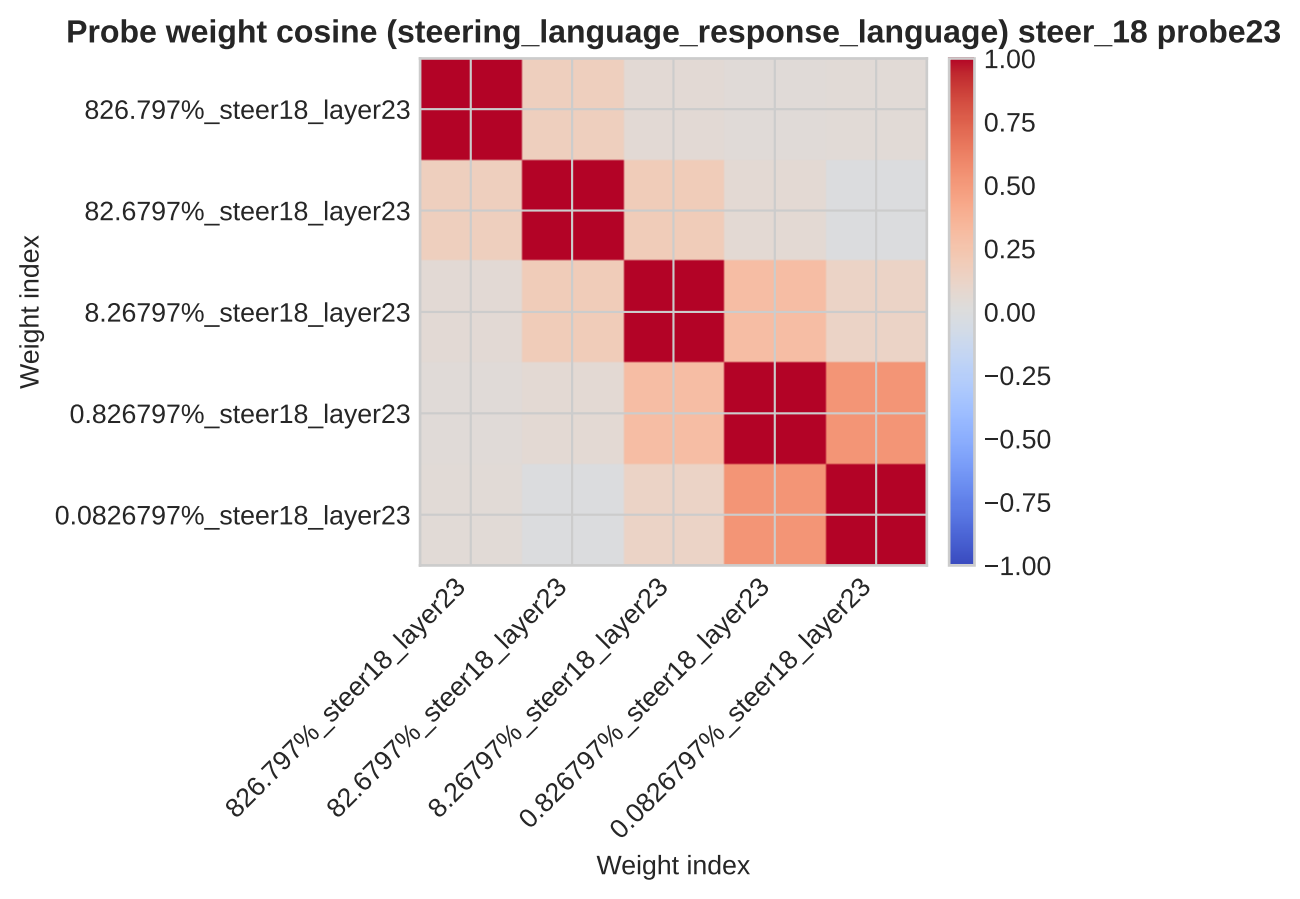
Probe weight cosine (steering_language_response_language) steer_18 probe23
826.797%_steer18_layer23
82.6797%_steer18_layer23
8.26797%_steer18_layer23
0.826797%_steer18_layer23
0.0826797%_steer18_layer23
Weight index
826.797%_steer18_layer23
82.6797%_steer18_layer23
8.26797%_steer18_layer23
0.826797%_steer18_layer23
0.0826797%_steer18_layer23
Weight index
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00