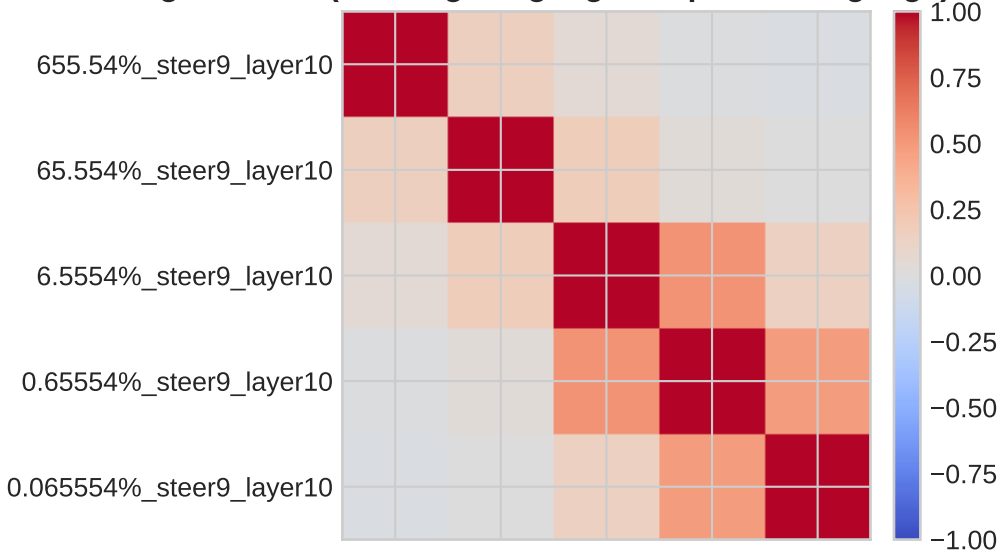


Probe weight cosine (steering\_language\_response\_language) steer\_9 probe10

Weight index



655.54%\_steer9\_layer10  
65.554%\_steer9\_layer10  
6.5554%\_steer9\_layer10  
0.65554%\_steer9\_layer10  
0.065554%\_steer9\_layer10

Weight index