

Probe weight cosine (steering\_language\_response\_language) steer\_18 probe19

