Probe weight cosine (steering_language_response_language) steer_8 probe8