Probe weight cosine (steering_language_response_language) steer_0 probe13