Probe weight cosine (steering_language_response_language) steer_17 probe17