Probe weight cosine (steering_detectable_format_json_format) steer_0 probe0