

Probe weight cosine (steering_language_response_language) steer_9 probe18

Weight index

