

Mean (and standard error) of Linear Probes Trained on ID Sets, Evaluated on On-Policy Sets Across All Behaviours

