

Probes Trained on ID Training Sets, Evaluated on ID and OOD On-Policy Test Sets, Averaged Across Behaviours

