

Attention Probes Trained on ID Training Sets, Evaluated on OOD On-Policy Test Sets for Each Behaviour

