
Attention Probes Trained on ID Training Sets, Evaluated on ID On-Policy Incentivised Test Sets for Each Behaviour
On-Policy Incentivised ID Test Set AUROC
1.0
0.9
0.8
0.7
0.6
0.5
Refusal
Lists
Metaphors
Science
Sycophancy_1
Sycophancy_2
Deferral_1
Deferral_2
Deception_1
Deception_2
Sandbagging_1
Sandbagging_2
Behaviour
ID Training Set
On-Policy Incentivised
On-Policy Prompted
Off-Policy