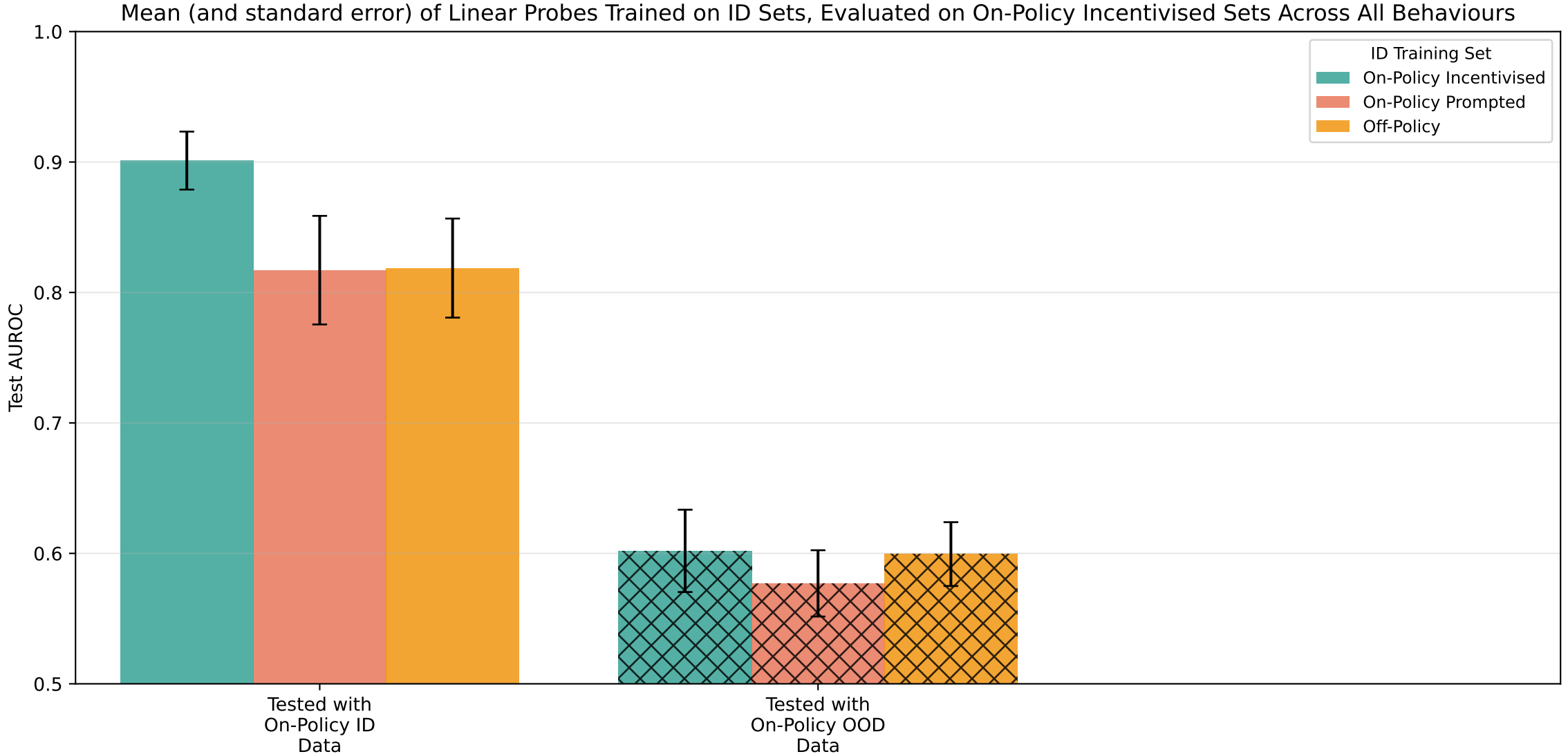
Mean (and standard error) of Linear Probes Trained on ID Sets, Evaluated on On-Policy Incentivised Sets Across All Behaviours
ID Training Set
On-Policy Incentivised
On-Policy Prompted
Off-Policy
Test AUROC
1.0
0.9
0.8
0.7
0.6
0.5
Tested with On-Policy ID Data
Tested with On-Policy OOD Data