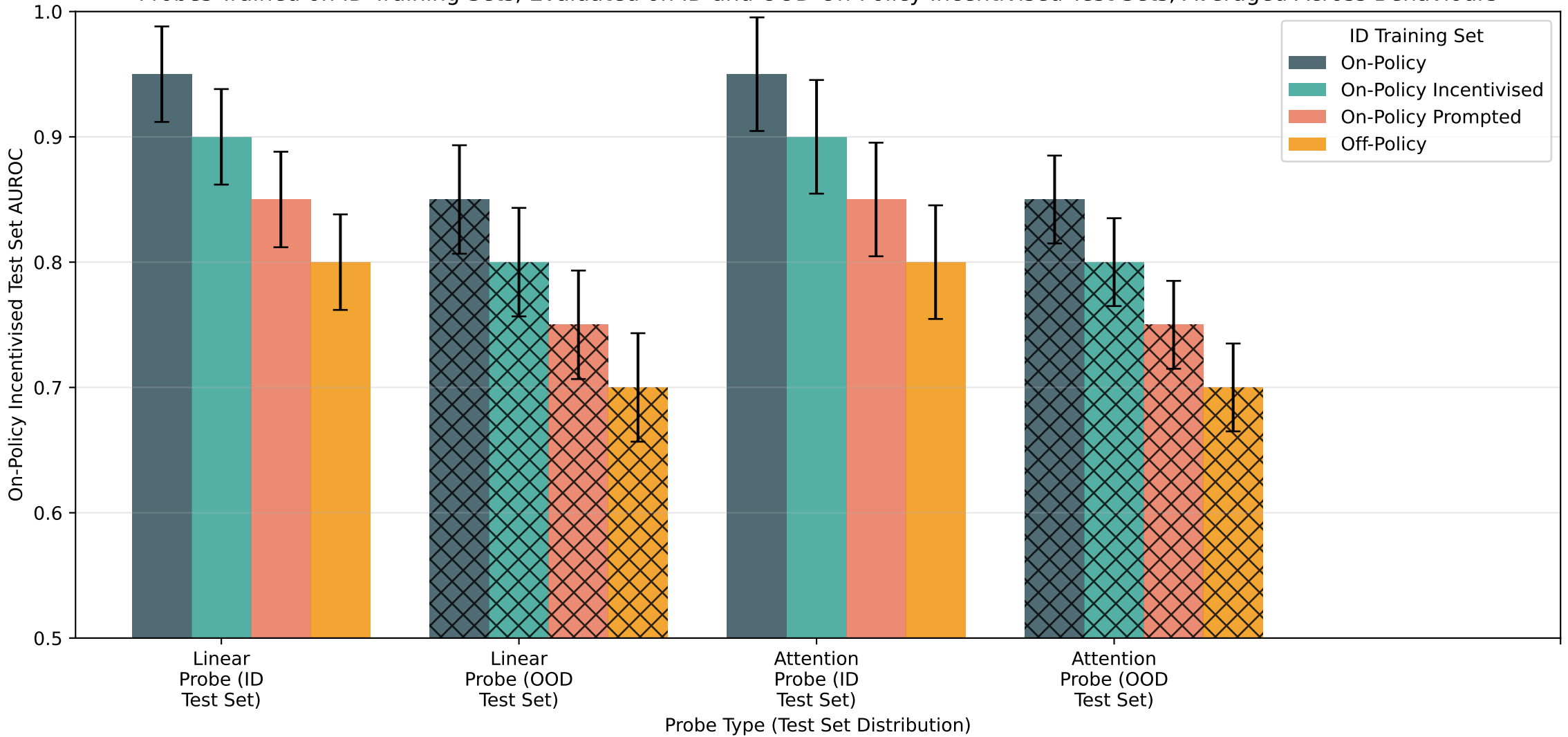
Probes Trained on ID Training Sets, Evaluated on ID and OOD On-Policy Incentivised Test Sets, Averaged Across Behaviours
On-Policy Incentivised Test Set AUROC
1.0
0.9
0.8
0.7
0.6
0.5
ID Training Set
On-Policy
On-Policy Incentivised
On-Policy Prompted
Off-Policy
Linear Probe (ID Test Set)
Linear Probe (OOD Test Set)
Attention Probe (ID Test Set)
Attention Probe (OOD Test Set)
Probe Type (Test Set Distribution)