

Original	0.71	0.71	0.71	0.68	0.71
GradDiff	0.28	0.65	0.53	0.31	0.41
NPO	0.27	0.65	0.50	0.28	0.41
SimNPO	0.28	0.64	0.55	0.30	0.50
RMU	0.27	0.70	0.41	0.26	0.59
RR	0.32	0.71	0.45	0.36	0.51
ELM	0.32	0.70	0.52	0.32	0.68
DPO	0.33	0.66	0.34	0.28	0.38
IDK+AP	0.32	0.52	0.40	0.29	0.29
NPO+SAM*	0.26	0.51	0.33	0.26	0.28
NPO+IRM*	0.27	0.56	0.34	0.26	0.30
RMU+LAT*	0.31	0.71	0.43	0.26	0.57
TAR*	0.27	0.68	0.33	0.25	0.29

Unlearned

RobReL

RobFT

RobQT

RobJA