

Original	0.70	0.69	0.66	0.58	0.70
GradDiff	0.03	0.31	0.34	0.43	0.54
NPO	0.02	0.23	0.26	0.41	0.53
SimNPO	0.00	0.10	0.06	0.42	0.60
RMU	0.16	0.35	0.28	0.38	0.68
RR	0.18	0.37	0.29	0.29	0.70
ELM	0.38	0.58	0.53	0.55	0.69
DPO	0.24	0.07	0.08	0.10	0.52
IDK+AP	0.07	0.40	0.32	0.32	0.42
NPO+SAM*	0.00	0.11	0.05	0.24	0.40
NPO+IRM*	0.00	0.06	0.02	0.09	0.47
RMU+LAT*	0.21	0.40	0.29	0.42	0.69
TAR*	0.00	0.06	0.03	0.10	0.50

Unlearned

RobFT (GSM8K)

RobFT (SST2)

RobFT (MNLI)

RobReL