

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

Unlearned  
*Rob<sub>FT</sub> (GSM8K)*   *Rob<sub>FT</sub> (SST2)*   *Rob<sub>FT</sub> (MNLI)*   *Rob<sub>ReL</sub>*