

Original	0.64	0.36	0.42	0.49	0.76
GradDiff	0.46	0.20	0.33	0.15	0.12
NPO	0.51	0.35	0.33	0.15	0.03
SimNPO	0.56	0.36	0.37	0.15	0.58
RMU	0.60	0.31	0.41	0.50	0.76
RR	0.63	0.36	0.42	0.49	0.76
ELM	0.62	0.33	0.42	0.48	0.73
DPO	0.55	0.28	0.42	0.54	0.75
IDK+AP	0.44	0.23	0.38	0.45	0.37
NPO+SAM*	0.51	0.24	0.30	0.16	0.03
NPO+IRM*	0.53	0.22	0.36	0.10	0.02
RMU+LAT*	0.60	0.30	0.41	0.46	0.73
TAR*	0.53	0.22	0.37	0.08	0.00

MMLU

TruthfulQA

MathQA

IFEval

GSM8K