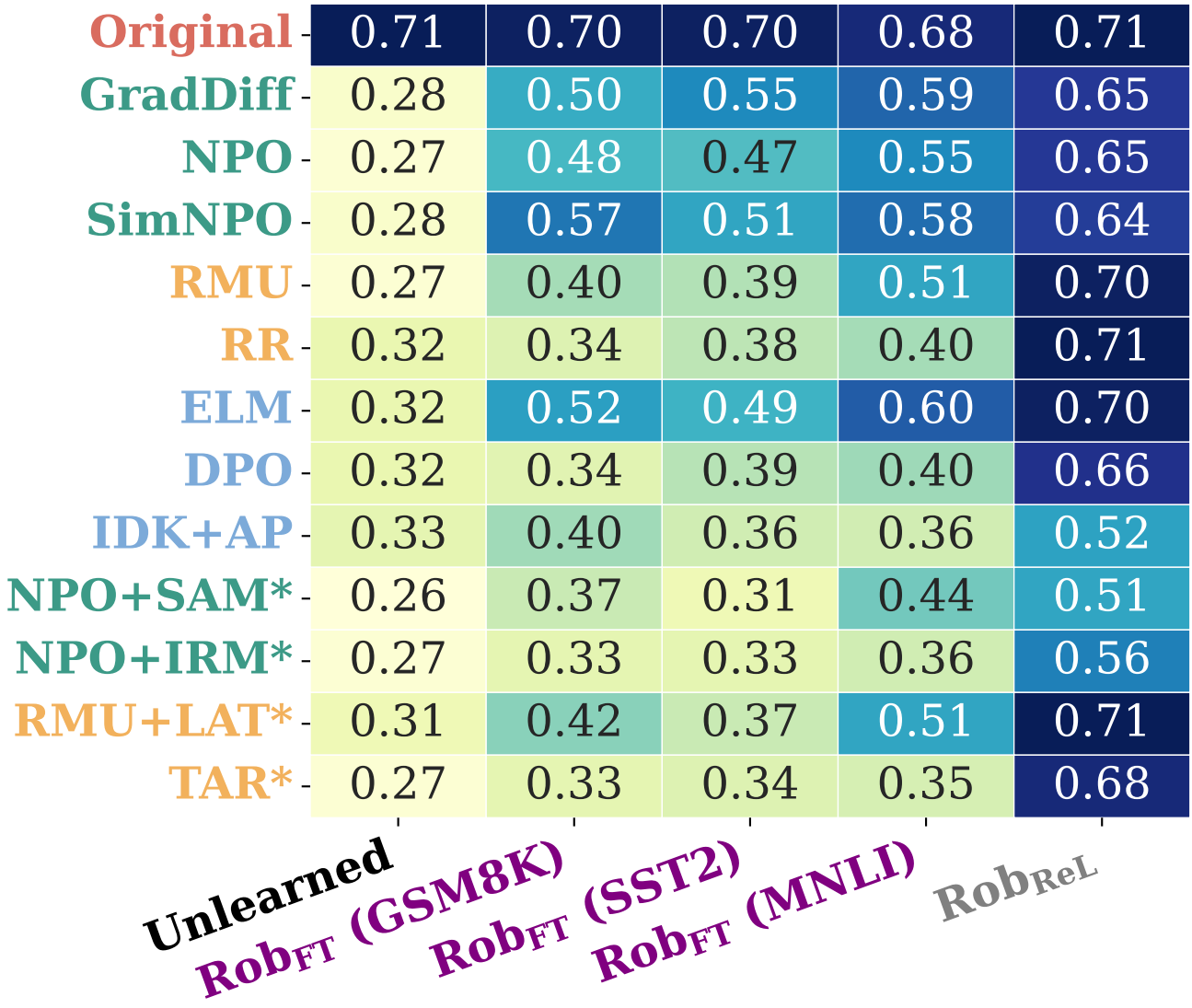
Original 0.71 0.70 0.70 0.68 0.71
GradDiff 0.28 0.50 0.55 0.59 0.65
NPO 0.27 0.48 0.47 0.55 0.65
SimNPO 0.28 0.57 0.51 0.58 0.64
RMU 0.27 0.40 0.39 0.51 0.70
RR 0.32 0.34 0.38 0.40 0.71
ELM 0.32 0.52 0.49 0.60 0.70
DPO 0.32 0.34 0.39 0.40 0.66
IDK+AP 0.33 0.40 0.36 0.36 0.52
NPO+SAM* 0.26 0.37 0.31 0.44 0.51
NPO+IRM* 0.27 0.33 0.33 0.36 0.56
RMU+LAT* 0.31 0.42 0.37 0.51 0.71
TAR* 0.27 0.33 0.34 0.35 0.68
Unlearned
Rob_FT (GSM8K)
Rob_FT (SST2)
Rob_FT (MNLI)
Rob_ReL